



**An investigation into the genetic architecture of  
multiple system atrophy and familial Parkinson's  
disease**

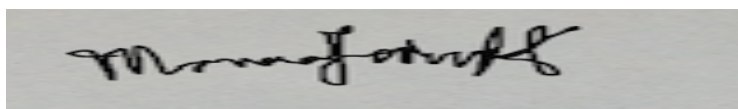
**By Monica Federoff**

A thesis submitted to University College London for the degree of  
Doctor of Philosophy

Laboratory of Neurogenetics, Department of Molecular  
Neuroscience, Institute of Neurology, University College London  
(UCL)

I, Monica Federoff, confirm that the work presented in this thesis is my own. Information derived from other sources and collaborative work have been indicated appropriately.

Signature:

A rectangular box containing a handwritten signature in black ink. The signature is cursive and appears to read "Monica Federoff".

Date: 09/06/2016

**Acknowledgements:**

When I first joined the Laboratory of Neurogenetics (LNG), NIA, NIH as a summer intern in 2008, I had minimal experience working in a laboratory and was both excited and anxious at the prospect of it. From my very first day, Dr. Andrew Singleton was incredibly welcoming and introduced me to my first mentor, Dr. Javier Simon-Sanchez. Within just ten weeks working in the lab, both Dr. Singleton and Dr. Simon-Sanchez taught me the fundamental skills in an encouraging and supportive environment. I quickly got to know others in the lab, some of whom are still here today, and I sincerely appreciate their help with my assimilation into the LNG.

After returning for an additional summer and one year as an IRTA postbac, I was honored to pursue a PhD in such an intellectually stimulating and comfortable environment. I am so grateful that Dr. Singleton has been such a wonderful mentor, as he is not only a brilliant scientist, but also extremely personable and approachable. If I inquire about meeting with him, he always manages to make time in his busy schedule and provides excellent guidance and mentorship.

I am extremely appreciative of the mentorship from my secondary supervisor, Dr. Henry Houlden, over the last three years. Despite my being a non-resident student, Dr. Houlden has been a great communicator and has been very helpful towards mapping out my future goals during my visits to UCL ION. As a physician-scientist, he has an excellent perspective on the balance required between both clinical and research aspects in one's career. I have also been afforded the opportunity to shadow him in the

Neurology clinic during my time at UCL, which has been invaluable towards connecting the fruits of neurogenetics research with patient care in clinical neurology.

Working with several other PhD students at UCL ION has been a pleasure. Lucia Schottlaender and Sarah Wiethoff have been wonderful colleagues and great friends. At the LNG, I have come to meet so many extraordinary people (from PhD students, to post docs and IRTAs) who I have both worked with and gotten to know including: Monia Hammer, Celeste Sassi, Christopher Letson, Timothy Ryan Price, Sampath Arepalli, Nick Bernstein, Sam Coster and many others. I would also like to specifically thank Ryan Price for helping me with the heritability analysis, as he has been a great asset to the LNG team. Secondly, I would like to greatly thank Dr. Mike Nalls and Jinhui Ding for helping me with my MSA exome sequencing analyses.

Finally, I would like to thank my wonderful family for their love and support during these last few years. I could not have done it without all of you by my side.

I'm so fortunate to have had this experience with everyone at LNG and UCL and look forward to future collaborations.



## Table of Contents

1	Introduction.....	18
1.1	Thesis objectives:.....	18
1.2	Why pursue genetics in disease? .....	20
1.3	Advances in genomic technologies: unraveling the genetic basis of disease .....	21
1.3.1	Monogenic diseases .....	22
1.3.1.1	Linkage analysis and positional cloning .....	23
1.3.1.2	Next generation sequencing in monogenic diseases .....	25
1.3.2	Complex diseases.....	30
1.3.2.1	Candidate gene association studies in complex diseases .....	33
1.3.2.2	Genome wide association studies in complex diseases .....	34
1.3.2.3	Next generation sequencing in complex diseases .....	44
1.3.3	The future of human disease genetics .....	48
1.4	Parkinson's disease.....	49
1.4.1	Clinical and neuropathological features of Parkinson's disease .....	49
1.4.2	Genetic etiology of monogenic PD .....	52
1.4.2.1	Autosomal dominant PD.....	53
1.4.2.1.1	LRRK2.....	53
1.4.2.1.2	SNCA.....	55
1.4.2.1.3	VPS35.....	57
1.4.2.1.4	ATXN2 and ATXN3 .....	57
1.4.2.1.5	MAPT .....	58
1.4.2.1.6	DCTN1.....	58
1.4.2.1.7	GCH-1.....	59
1.4.2.2	Autosomal recessive PD .....	59
1.4.2.2.1	PARK2.....	59
1.4.2.2.2	PINK1 .....	60
1.4.2.2.3	DJ-1/PARK7 .....	60
1.4.2.2.4	ATP13A2 .....	61
1.4.2.2.5	FBXO7.....	62
1.4.2.2.6	PLA2G6.....	62
1.4.2.2.7	PANK2.....	63
1.4.2.2.8	Other rare recessive forms of PD.....	64
1.4.3	Molecular mechanisms of PD gene mutations.....	65
1.4.4	Integrating critical molecular processes regulated by PD proteins.....	66
1.4.5	Risk loci in PD .....	69
1.4.5.1	Risk loci identified by GWA .....	71
1.4.6	Interpretation of GWA findings and PD etiology .....	75
1.4.7	Making progress in PD .....	76
1.5	Multiple system atrophy .....	77
1.5.1	Clinical and neuropathological features of MSA.....	78
1.5.2	Understanding MSA etiology .....	83
1.5.3	Preliminary association studies.....	83
1.5.4	The genetics of MSA .....	85
1.5.4.1	Mendelian inheritance.....	85

1.5.4.2	COQ2 mutations .....	86
1.5.4.3	Genes encoding proteins involved in oxidative stress .....	88
1.5.4.4	PRNP.....	89
1.5.4.5	SNCA.....	90
1.5.4.6	Other PD linked genes .....	92
1.5.4.7	Copy number changes.....	94
1.5.5	Proposed mechanisms of MSA pathogenesis .....	96
1.5.5.1	Role of Neurotoxicity and Oxidative Stress .....	96
1.5.5.2	Role of ubiquitin-proteasome system .....	100
1.5.6	The dynamic behind key players: neurotoxicity, oxidative stress and the UPS 103	
1.5.7	Drug Therapies and targets in MSA .....	104
1.5.8	A comparison of PD and MSA .....	106
1.5.9	How to move forward .....	107
2	Estimating the heritable component of MSA.....	110
2.1	Introduction.....	110
2.2	Materials and methods .....	112
2.2.1	Subjects .....	112
2.2.2	Pre-imputation base calling quality control .....	113
2.2.3	Imputation .....	114
2.2.4	Genome-wide complex trait analysis .....	115
2.2.5	Bayesian estimate of PD-derived heritability .....	115
2.3	Results.....	116
2.3.1	Quality Control .....	116
2.3.2	Post-imputation GWA .....	117
2.3.3	Post-imputation candidate gene analysis .....	118
2.3.4	Heritability analysis .....	119
2.3.5	Bayesian estimate of PD-derived heritability .....	127
2.4	Discussion.....	129
3	Identifying candidate genes and variants for MSA using exome sequencing .....	138
3.1	Introduction.....	138
3.2	Materials and methods .....	140
3.2.1	Subjects .....	140
3.2.2	Whole exome sequencing .....	143
3.2.2.1	DNA library prep & enrichment .....	144
3.2.2.2	Cluster generation .....	144
3.2.2.3	Parallel sequencing by synthesis.....	145
3.2.2.4	Data analysis .....	145
3.2.3	Illumina TruSeq protocol.....	146
3.2.3.1	DNA Library preparation and enrichment .....	146
3.2.3.1.1	Quantification .....	146
3.2.3.1.2	Fragmentation of gDNA .....	146
3.2.3.1.3	Quality check using the Bioanalyzer .....	146
3.2.3.1.4	Post fragmentation end repair .....	150
3.2.3.1.5	Cleaning with AMPure Beads XP .....	150
3.2.3.1.6	3' End Adenylation.....	151

3.2.3.1.7	Adapter Ligation .....	151
3.2.3.1.8	DNA library enrichment .....	152
3.2.3.1.9	Exome Capture .....	152
3.2.3.1.10	First Hybridization .....	153
3.2.3.1.11	First Wash .....	153
3.2.3.1.12	Second Hybridization .....	154
3.2.3.1.13	Second Wash .....	155
3.2.3.1.14	Library Enrichment .....	155
3.2.4	Illumina Nextera rapid capture protocol .....	156
3.2.4.1	DNA Library preparation and enrichment .....	156
3.2.4.1.1	Quantification .....	156
3.2.4.1.2	Tagmentation of gDNA .....	158
3.2.4.1.3	Clean up Tagmented DNA .....	158
3.2.4.1.4	First PCR Amplification .....	159
3.2.4.1.5	First PCR clean up .....	160
3.2.4.1.6	Quality check using the Bioanalyzer .....	160
3.2.4.1.7	First Hybridization .....	160
3.2.4.1.8	First Capture .....	161
3.2.4.1.9	First Elution .....	161
3.2.4.1.10	Second Hybridization .....	162
3.2.4.1.11	Second Capture .....	162
3.2.4.1.12	Capture sample clean up .....	162
3.2.4.1.13	Second PCR Amplification .....	162
3.2.4.1.14	Second PCR clean up .....	163
3.2.4.2	DNA amplification and clustering on the C-Bot .....	163
3.2.4.3	Parallel sequencing by synthesis on the Illumina Hi Seq 2000 .....	164
3.2.5	Raw data analysis .....	166
3.2.5.1	Mapping, alignment and duplicate removal .....	166
3.2.5.2	Raw variant callings and file conversions .....	167
3.2.5.3	Incorporation of reference databases .....	169
3.2.5.4	Assignment of quality scores (Phredd scores) to all variant calls .....	170
3.2.5.5	Generation of variant call files (VCF) and (group) gVCFs .....	170
3.2.5.6	Downstream analysis and filtering of VCFs .....	170
3.2.6	Looking for variants in PD risk and causal genes .....	171
3.2.7	Variant and Gene based Approach Filtering Pipelines .....	171
3.2.8	Annovar Filtering Process .....	173
3.2.9	Sanger sequencing confirmation of variants .....	174
3.2.10	Individual Variant and Gene Burden Testing .....	177
3.2.10.1	RAREMETAL Analysis: Quality Control .....	179
3.2.10.2	RAREMETAL analysis: gene burden and single variant testing .....	180
3.2.10.2.1	Single variant analysis: variance component (non-burden) tests .....	181
3.2.10.2.2	Weighted Aggregation Test .....	181
3.2.10.2.3	Adaptive Burden test .....	182
3.2.10.2.4	Combined Burden test .....	182
3.3	Results .....	183
3.3.1	Quality control filtering of locally executed analyses .....	183

3.3.2	Looking for variants in PD associated genes .....	185
3.3.3	Filtering through a Variant Based Approach .....	186
3.3.4	Filtering through a Gene Based Approach.....	188
3.3.5	RAREMETAL Individual Variant and Gene Burden Analyses .....	193
3.3.5.1	Quality Control with GoogleGenome data in GoogleCloud .....	193
3.3.5.2	Gene Burden Results.....	196
3.3.5.2.1	In-depth gene analysis .....	198
3.3.5.3	Single Variant Results.....	203
3.3.5.3.1	Overview of Results.....	203
3.3.5.3.2	Comparison with MSA GWAS results .....	210
3.3.5.3.3	Comparison with WES results .....	210
3.4	Discussion.....	211
4	Exploring the genetic etiology of PD in the Greek village of Rapsani.....	222
4.1	Introduction.....	222
4.2	Materials and methods .....	225
4.2.1	Subjects .....	225
4.2.2	Genotyping.....	229
4.2.3	Quality control .....	231
4.2.4	Identifying runs of homozygosity .....	231
4.2.5	Identifying segments identical by descent .....	232
4.2.6	Whole exome sequencing .....	233
4.2.7	Whole genome sequencing .....	235
4.2.8	<i>C9ORF72</i> hexanucleotide repeat screening.....	238
4.3	Results.....	240
4.3.1	Ancestral background .....	240
4.3.2	Genotyping.....	241
4.3.3	Whole exome sequencing .....	248
4.3.3.1	Quality control filtering .....	248
4.3.3.2	Examining known PD genes.....	249
4.3.3.3	Population database filtering.....	249
4.3.4	Whole genome sequencing .....	252
4.3.4.1	Quality Control Filtering.....	252
4.3.4.2	Homozygosity Mapper Analyses .....	253
4.3.4.2.1	Structural variant homozygosity results .....	253
4.3.4.2.2	SNP indel homozygosity results .....	254
4.3.4.3	Whole genome sequencing variant filtering pipeline analysis .....	255
4.3.4.3.1	Copy number variant data analysis.....	256
4.3.4.3.2	Structural variant data analysis .....	256
4.3.4.3.3	SNP Indel analysis .....	256
4.3.5	<i>C9ORF72</i> screening.....	258
4.4	Discussion.....	259
5	Conclusions and future directions.....	266
5.1.1	Chapter 2 overview .....	267
5.1.2	Chapter 3 overview .....	268
5.1.3	Chapter 4 overview .....	270
5.1.4	Final thoughts and future directions.....	272

6	Acknowledgements.....	274
7	References.....	275
8	Appendix.....	315
	8.1.1 Transcripts.....	315
	8.1.1.1 MSA local pipeline: WES variant filtering approach.....	315
	8.1.1.2 MSA local pipeline: WES gene filtering approach.....	315
	8.1.1.3 MSA Googlegenome pipeline: Gene burden and single variant analyses	316
	8.1.1.4 Greek Rapsani PD WES candidates .....	316
	8.1.1.5 Greek Rapsani PD WGS candidates .....	317
	8.1.2 Primer sequences .....	317
	8.1.2.1 MSA WES variant filtering approach candidates.....	317
	8.1.2.2 MSA WES gene filtering approach candidates .....	318
	8.1.2.3 MSA Googlegenome pipeline: Gene burden and single variant analyses	318
	8.1.2.4 Greek Rapsani PD WES candidates .....	319
	8.1.2.5 Greek Rapsani PD WGS candidates .....	319
	8.1.3 Cyclor programs.....	320
	8.1.4 PCR master mixes .....	321
	8.1.5 MSA GWA study results .....	321

## Abbreviations

AD	Alzheimer's Disease
ADGC	Alzheimer's Disease Genetic Consortium
ALS	Amyotrophic Lateral Sclerosis
CBD	Corticobasal Degeneration
CDCV	Common Disease Common Variant
CDRV	Common Disease Rare Variant
CHARGE	Cohorts for Heart and Aging Research in Genomic Epidemiology
CNV	Copy Number Variant
DLB	Dementia with Lewy Bodies
DNA	Deoxyribonucleic Acid
dsDNA	double stranded DNA
DMSO	Dimethyl Sulfoxide
ENCODE	Encyclopedia of DNA Elements
EVS	Exome Variant Server
FALS	Familial Amyotrophic Lateral Sclerosis
FC	Flow Cell
FTD	Frontotemporal Dementia
GCI	Glial Cytoplasmic Inclusion
GCTA	Genome-wide Complex Trait Analysis
gDNA	Genomic DNA
GWA	Genome Wide Association
HWE	Hardy-Weinberg Equilibrium
Indels	Insertion/Deletion Variants
KTS	Kohlschutter-Tonz Syndrome
LB	Lewy Body
LD	Linkage Disequilibrium
MAF	Minor Allele Frequency
MRI	Magnetic Resonance Imaging
MRV	Multiple Rare Variant
MSA	Multiple System Atrophy
NGS	Next Generation Sequencing
OR	Odds Ratio
PCR	Polymerase Chain Reaction
PD	Parkinson's Disease
PSP	Progressive Supranuclear Palsy
QC	Quality Control
REML	Restricted Maximum Likelihood
RFLP	Restriction Fragment Length Polymorphism
RNA	Ribonucleic Acid
RSB	Resuspension Buffer
SNP	Single Nucleotide Polymorphism
SNPC	Substantia Nigra Pars Compacta
SPECT	Single Photon Emission Computerized Tomography
UPS	Ubiquitin proteasome system

UTR	Untranslated Region
VIF	Variance Inflation Factor
WES	Whole Exome Sequencing
WGS	Whole Genome Sequencing

## List of Tables

Table 1: Neuropathology of monogenic forms of PD. ....	51
Table 2: Results of PD GWA study.....	73
Table 3: Summary statistics of samples included in GCTA analysis. ....	117
Table 4: Results of 20kb windows between PD initiation and termination of PD GWA hits.....	119
Table 5: Heritability estimate by cohort and subgroup.....	122
Table 6: Clinical cohorts.....	123
Table 7: Pathologically confirmed cohorts.....	124
Table 8: Heritability estimates by chromosome. ....	126
Table 9: Descriptive statistics of MSA WES cohort.....	140
Table 10: Origin of samples by contributing center.....	142
Table 11: Results of variants among all MSA samples in PD associated genes. ....	186
Table 12: Sanger sequencing results using variant-based approach for MSA WES analysis.....	187
Table 13: Sanger sequencing results based on gene-based approach analysis. ....	191
Table 14: All 27 sanger sequencing confirmed variants from WES with Mutation Taster prediction. ....	193
Table 15: The most significant genes with p-values $< 1 \times 10^{-6}$ among all 3 gene burden tests (MB, VT, CMC). ....	197
Table 16: Non-synonymous <i>LRRK2</i> variants identified by gene burden analyses. ....	198
Table 17: Non-synonymous <i>PARK2</i> variants identified by gene burden analyses.....	201
Table 18: Non-synonymous <i>EIF4G1</i> variants identified by gene burden analyses.....	201
Table 19: Non-synonymous <i>GIGYF2</i> variants identified by gene burden analyses.....	201
Table 20: Non-synonymous <i>VPS13D</i> variants identified by gene burden analyses.....	202
Table 21: Non-synonymous <i>SLC44A5</i> variants identified by gene burden analyses. ....	202
Table 22: Non-synonymous <i>GLIPR1</i> variants identified by gene burden analyses. ....	203
Table 23: Non-synonymous <i>CASP8AP2</i> variants identified by gene burden analyses. .	203
Table 24: A list of coding, protein-altering, highly significant single variants with p-values $< 0.05$ .....	209
Table 25: Single non-synonymous variants with p-values $< 0.004$ in top genes from MSA GWA study .....	210
Table 26: Significant non-synonymous single variants in genes identified by WES filtering pipelines .....	211
Table 27: Most significant and functionally relevant genes identified in hypothesis generating dataset.....	216
Table 28: Clinical Phenotypes of Rapsani families I-III. ....	228
Table 29: PCR master mix used for <i>C9ORF72</i> screening in all Greek samples .....	239
Table 30: Repeat primer mix used for <i>C9ORF72</i> screening in all Greek samples.....	239
Table 31: Observed and expected rates of homozygosity and inbreeding coefficients of Rapsani villagers.....	241
Table 32: Quality control of Greek PD Rapsani cohort genotyping.....	242
Table 33: Homozygosity mapper gene distiller output of Greek Rapsani cohort.....	245
Table 34: Results of “Shared Segment analysis” using Plink and GERMLINE. ....	247



Table 35: All PD-associated polymorphisms identified in 23 members of Greek Rapsani cohort. ....	249
Table 36: WES sanger sequencing results for individual Rapsani families .....	251
Table 37: Sanger sequencing results of variants identified using a WGS liberal filtering approach .....	258
Table 38: PD genes in MSA VCF that did not confirm with Sanger sequencing .....	315
Table 39: Variants predicted very damaging in MSA VCF not confirmed with Sanger sequencing.....	315
Table 40: All variants in MSA VCF confirmed with Sanger sequencing .....	315
Table 41: All variants that did not confirm with Sanger sequencing .....	316
Table 42: All variants investigated in the “in-depth gene” analysis. ....	316
Table 43: Variants checked with Sanger sequencing in Greek PD cohort from WES data .....	316
Table 44: Variants checked with Sanger sequencing in Greek PD individual families from WES data .....	316
Table 45: Variants checked with Sanger sequencing in Greek PD cohort from WGS data .....	317
Table 46: PD genes in MSA VCF that did not confirm with Sanger sequencing .....	317
Table 47: Variants predicted very damaging in MSA VCF not confirmed with Sanger sequencing.....	317
Table 48: All variants in MSA VCF confirmed with Sanger sequencing .....	318
Table 49: All variants that did not confirm with Sanger sequencing .....	318
Table 50: Variants checked with Sanger sequencing in Greek PD cohort from WES data .....	319
Table 51: Variants checked with Sanger sequencing in Greek PD individual families from WES data .....	319
Table 52: Variants checked with Sanger sequencing in Greek PD cohort from WGS data .....	319
Table 53: 72 touchdown 56 PCR cyclers conditions used for all primers. Total time: 2h, 45min .....	320
Table 54: Sequencing cyclers conditions used for all primers. Total time: 2h, 22min ....	320
Table 55: PCR Mastermixes tested and utilized for all primers .....	321
Table 56: MSA GWA Study results.....	322

## List of Figures

Figure 1: The genetic basis of disease. ....	21
Figure 2: Types of Mendelian inherited disorders. ....	22
Figure 3: Approximate number of gene discoveries made by WES and WGS versus conventional approaches since 2010. ....	27
Figure 4: Whole exome sequencing analysis schema. ....	28
Figure 5: Gene discovery methods. ....	33
Figure 6: A schematic of how imputation works. ....	35
Figure 7: Number of GWA studies published per year. ....	37
Figure 8: Direct and indirect nature of associated variants. ....	40
Figure 9: Microscopic findings in PD. ....	50
Figure 10: Putative molecular mechanisms underlying PD. ....	68
Figure 11: Shared and distinguishing pathogenic, pathologic and clinical features of MSA-P and PD. ....	106
Figure 12: Principal components of MSA samples. ....	114
Figure 13: Manhattan plot of post-imputation MSA GWA study. ....	118
Figure 14: Heritability by cohort in diagnostic subgroups ....	121
Figure 15: Chromosomal heritability estimates by diagnostic subgroup. ....	127
Figure 16: Disease-specific heritability estimates. ....	130
Figure 17: Origin of MSA cohort samples ....	141
Figure 18: Overview of WES pipeline. ....	143
Figure 19: High quality library sample on the Agilent bioanalyzer. ....	150
Figure 20: Electropherogram of a successful library. ....	155
Figure 21: Nextera rapid capture enrichment process. ....	157
Figure 22: Tagmentation followed by first PCR. ....	158
Figure 23: C-Bot clustering showing bridge amplification and DNA cluster formation. .....	164
Figure 24: Flowcell with amplified DNA clusters. ....	164
Figure 25: Parallel sequencing by synthesis on the Illumina Hi Seq 2000. ....	165
Figure 26: Whole exome sequencing analysis pipeline. ....	167
Figure 27: Bioinformatics Pipeline of Raw Variant Callings and File Conversions. ....	168
Figure 28: Read coverage of <i>APP</i> using whole exome sequencing and whole genome sequencing. ....	169
Figure 29: Public Reference Databases used as exclusion criteria for WES analysis. ....	171
Figure 30: Variant Filtering Pipeline used for MSA whole exome sequencing analysis. .....	173
Figure 31: Sanger sequencing PCR cleanup using Ampure paramagnetic beads. ....	176
Figure 32: Merging clean MSA sample cohort with 1300 control samples to make final VCF. ....	177
Figure 33: Pipeline used for data pooling on googleCloud before running RAREMETAL analyses. ....	179
Figure 34: 10X and 30X depth of 411 MSA exome sequenced samples. ....	183
Figure 35: Mean depth per individual in MSA samples (top) and 1300 controls (bottom) .....	184

Figure 36: Multidimensional scaling of 411 MSA samples to identify and remove population outliers .....	185
Figure 37: IGV example of high quality depth and coverage in candidate gene, <i>MAGEL2</i> .....	189
Figure 38: Sanger sequencing results in MSA sample heterozygous for <i>CASP8AP2</i> , p.M668T.....	192
Figure 39: Sanger sequencing results in MSA sample heterozygous for <i>RGS11</i> , p.G791A .....	192
Figure 40: Population stratification of MSA cases and controls .....	194
Figure 41: QQ plot of single variant results with a $MAF < 0.01$ .....	195
Figure 42: QQ plot of CMC burden test results with a $MAF < 0.01$ .....	196
Figure 43: Manhattan plot of VT gene burden test results in coding alleles with a $MAF < 0.01$ .....	197
Figure 44: MDS plot of MSA samples with 10 individuals carrying LRRK2 p.G2385R in coral.....	200
Figure 45: Ensembl Variant Effect Predictor Tool results of individual variant consequences.....	204
Figure 46: Manhattan plot of top hits from single variant analyses in MSA exome cohort. ....	205
Figure 47: Filtering results of single variant burden analyses for MSA exome cohort..	206
Figure 48: Rapsani, Greece.....	223
Figure 49: Pedigree of Rapsani families I and III.....	226
Figure 50: Pedigree of Rapsani family II.....	227
Figure 51: Pedigrees of Rapsani families IV and V .....	228
Figure 52: Illumina Infinium Genotyping Workflow.....	229
Figure 53: Overall genotyping workflow .....	230
Figure 54: Whole exome sequencing variant filtering pipeline for Greek Rapsani PD cohort .....	235
Figure 55: Macrogen whole genome sequencing analytical workflow .....	236
Figure 56: Whole genome sequencing variant filtering pipeline for Greek Rapsani PD cohort .....	238
Figure 57: Multidimensional Scaling of Greek Rapsani village members.....	240
Figure 58: Greek Rapsani cohort genotyping data viewed in homozygosity mapper ....	242
Figure 59: Greek Rapsani cohort genotyping data viewed in homozygosity mapper. ...	243
Figure 60: An example of a region of homozygosity along chromosome 6 for all Rapsani individuals.....	244
Figure 61: Example of artifact of novel SLC44A4 variant based on BAMs.....	246
Figure 62: Whole exome sequencing 10X and 30X depth of Greek Rapsani village samples.....	248
Figure 63: Whole exome sequencing: mean depth per individual in 22 Greek Rapsani PD samples.....	248
Figure 64: Greek Rapsani PD whole genome sequencing quality control results.....	253
Figure 65: Greek Rapsani cohort WGS SV data viewed in homozygosity mapper .....	254
Figure 66: Greek Rapsani cohort WGS SNP indel data viewed in homozygosity mapper .....	255

Figure 67: Example of positive and negative control for <i>C9ORF72</i> hexanucleotide repeat screening. ....	258
--------------------------------------------------------------------------------------------------------------	-----

**Work from this thesis has been published in the following articles**

- 1) Anna Sailer, MD, PhD, Sonja W. Scholz, MD, PhD, Michael A. Nalls, PhD, Claudia Schulte, PhD, Monica Federoff, MS, T. Ryan Price, MS, Andrew Lees, MD, Owen A. Ross PhD, Dennis W. Dickson, MD, Kin Mok, PhD, Niccolo E. Mencacci, MD, Lucia Schottlaender, MD, Viorica Chelban, MD, Helen Ling, PhD, Sean S. O'Sullivan, MD, PhD, Nicholas W. Wood, MD, PhD, Bryan J. Traynor, MD, PhD, Luigi Ferrucci, MD, PhD, Howard J. Federoff, MD, PhD, Timothy R. Mhyre, PhD, Huw R. Morris, PhD, Günther Deuschl, MD, Niall Quinn, MD, Hakan Widner, MD, PhD, Alberto Albanese, MD, Jon Infante, MD, Kailash P. Bhatia, MD, Werner Poewe, MD, Wolfgang Oertel, MD, Ullrich Wüllner, MD, Stefano Goldwurm, MD, PhD, Maria Teresa Pellecchia, MD, Joaquim Ferreira, MD, Eduardo Tolosa, MD, Bastian Bloem, MD, Olivier Rascol, MD, Wassilios G. Meissner, MD, PhD, John A. Hardy, PhD, MD, Tamas Revesz, MD, Janice L. Holton, MD, PhD, Thomas Gasser, MD, Gregor K. Wenning, MD, PhD, Andrew B. Singleton, PhD, Henry Houlden, MD, PhD, On behalf of the European Multiple System Atrophy Study Group and the UK Multiple System Atrophy Study Group. A Genome-Wide Association Study in Multiple System Atrophy. *Neurology*. 2016 Oct; 87(15):1591-1598.
- 2) Federoff M, Price TR, Sailer A, Scholz S, Hernandez D, Nicolas A, Singleton AB, Nalls M, Houlden H. Genome-wide estimate of the heritability of Multiple System Atrophy. *Parkinsonism Relat Disord*. 2016 Jan; 22:35-41
- 3) Federoff M, Schottlaender LV, Houlden H, Singleton A. Multiple system atrophy: the application of genetics in understanding etiology. *Clinical Autonomic Research*. 2015 Feb; 25(1):19-36. doi: 10.1007/s10286-014-0267-5.

# 1 Introduction

## 1.1 Thesis objectives:

Within the past decade, significant genetic underpinnings of devastating neurological disorders including Parkinson's disease (PD), Alzheimer's disease (AD) and Amyotrophic Lateral Sclerosis (ALS) have all been illuminated through advanced genomic technologies. Such discoveries have been facilitated by genome wide association (GWA) studies and next generation sequencing (NGS) investigations in order to identify common variants and rare variants, respectively, which contribute to disease risk. This thesis aims to extend these analyses to an understudied disease, multiple system atrophy (MSA), and to investigate the genetic basis of an apparent cluster of PD cases in Greece. Thus, I plan to accomplish three main goals in my thesis:

First, I would like to determine if common variants are associated with MSA risk through heritability analysis and if so, can these be identified through imputation of GWA study data using greater than 900 sporadic MSA cases. While common variants harboring an association with MSA may be either protective or deleterious, any significant findings will yield insight into the pathogenesis of disease.

Secondly, I would to identify candidate variants and gene-based variability that are associated with MSA using next generation sequencing in approximately 415 samples, about half of which are pathologically confirmed. Ideally, these will be putative causal variants that will shed light on the molecular mechanisms of disease and potential for therapeutic design in the future. However, it is feasible that such rare variants may modulate risk for MSA development. Because MSA is a rare disease systematic

investigation of the genetic basis of this disease has been challenging; the cohorts studied are small and thus lack power. Hence, the goal of this work is to produce rational evidence based candidate genes and variants for validation and replication by the MSA research community. This will be a cardinal step towards our understanding of the genetic basis of this severely debilitating and fatal neurodegenerative disorder.

Thirdly, I would like to further explore the genetic architecture of Parkinson's disease among a large Greek kindred that we believe has maintained a high degree of genetic isolation for the last several centuries. While we have identified several risk factors and causal variants associated with Parkinson's disease, heritability estimates suggest that other genetic variants remain to be found. By utilizing some of the most advanced technological approaches in genetics, I hope to elucidate a missing piece of the puzzle in the etiology and molecular underpinnings of this devastating disease.

## **1.2 Why pursue genetics in disease?**

To pursue the study of genetics, one is usually motivated by an interest in anthropological, scientific or medical questions. While the study of human anthropology has been critical towards discovering our ancestral origins as species, the drive to understand human physiology and pathophysiology, on both a molecular and gross level, is believed to be a critical milestone in the progress of modern medicine toward etiologic based therapies.

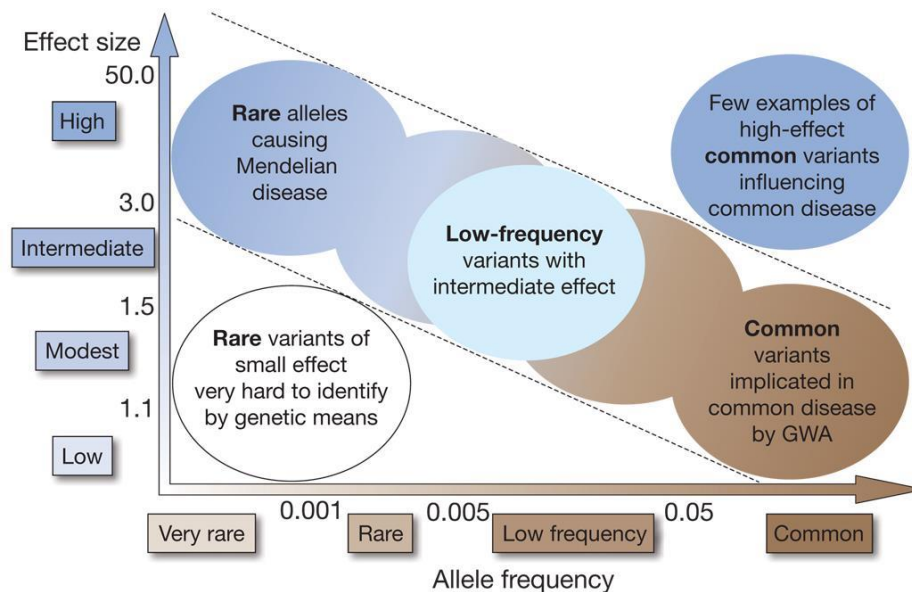
It is believed that through genetics, we can understand the molecular basis of disease, and that this understanding will afford the opportunity to develop and test etiologic based therapies. As we continue to unravel both risk factors and causes of disease through genetic analyses, we make significant strides in diagnosis and treatment, with the goal of seeking more preventative avenues in future medicine. As no individual is immune to the effects and implications of genetics, the scientific pursuit of genetics is essential for the continued prosperity of human health and vitality.

Throughout my journey as a PhD student, I have developed a profound appreciation for the current state of the field of genetics. This not only entails the remarkable progress in recent history but the passion and drive that is so inspiring from the scientific community. While we must remember that the number of failures will greatly outweigh our successes, the journey will be valuable both personally and for the greater scientific community, as we learn which areas to draw our attention towards or away from. Finally, following countless efforts of failure, we must believe they will make future success in our genetic studies that much more gratifying for scientific and medical communities alike.



### 1.3 Advances in genomic technologies: unraveling the genetic basis of disease

The salience of human genetics for the human condition has reached new heights and heralds the redefinition of clinical nosology. The intellectual ambition of preeminent scientists to collaborate in order to harness technological advances have led to three fundamental paradigms explaining the genetic etiology of human disease as depicted in Figure 1<sup>1</sup>.



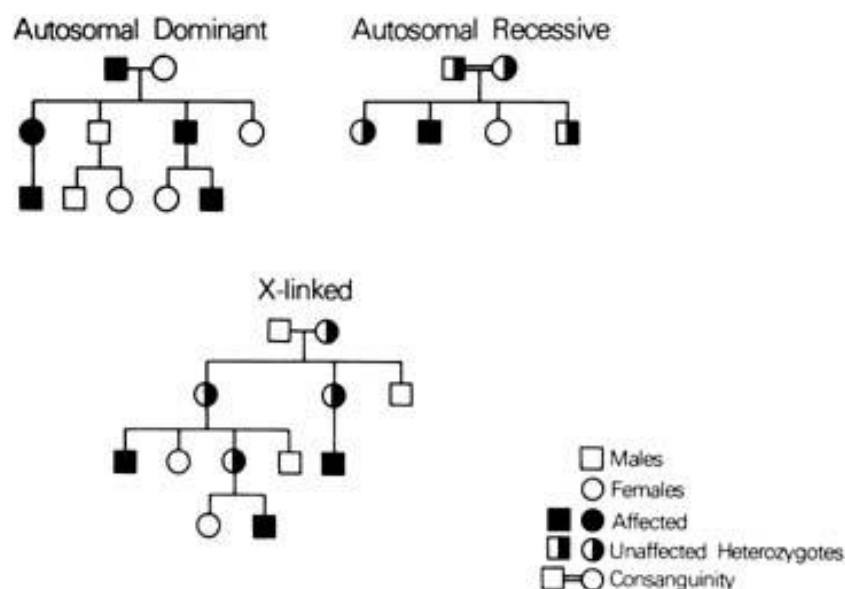
**Figure 1: The genetic basis of disease.**

Graphical depiction of common, low-frequency and rare variants and their corresponding effect size on association with disease. The top left corner includes very rare variants with large effect sizes, such as *SNCA* duplication or triplication or *LRRK2* p.G2019S in PD. An example of low-frequency variants with intermediate effects in PD would be heterozygous alleles in *GBA* that increase PD risk more than 5-fold. Common variants in common disease ascertained by GWA studies would be any of the replicated and validated PD GWA “hits,” such as variants in *STK39* or *HLA-DQB1*. A classic example of a high-effect common variant influencing common disease would be *APOE* in Alzheimer’s Disease. Finally, rare variants with small effects that are difficult to detect are largely unknown in the etiology of many complexes diseases.

Reproduced from (Manolio et al., 2009).<sup>1</sup>

### 1.3.1 Monogenic diseases

Perhaps the simplest category of genetic influence in disease is that of monogenic disorders, or “Mendelian inheritance”. This centers on disease causing genetic variability that is always, or highly likely to be, causal (highly or fully-penetrant diseases). These diseases may be inherited in an autosomal dominant, autosomal recessive or X-linked fashion (Figure 2).



**Figure 2: Types of Mendelian inherited disorders.**

Pedigrees depicting each mode of inheritance. Autosomal dominant diseases affect each successive generation while those of autosomal recessive can “skip” generations. X-linked diseases are carried by females in a heterozygous state and are described as having a carrier status; however, as males only have a single X-chromosome, variants are hemizygous, typically with a fully penetrant phenotype.

Reproduced from [http://resources.ama.uk.com/glowm\\_www/graphics/figures/v3/1150/001f.jpg](http://resources.ama.uk.com/glowm_www/graphics/figures/v3/1150/001f.jpg)

Pathological mutations, in the form of protein coding variants, copy number variations (CNVs), copy neutral variations (translocations or inversions) and expanded repetitive sequences, have been shown to cause monogenic disorders<sup>2,3</sup>. While such mutations are generally rare, there has been a great deal of success in identifying this

form of genetic influence in disease; further, such mutations have served as the basis for the majority of investigation into the molecular mechanisms that represent the disease process.

#### ***1.3.1.1 Linkage analysis and positional cloning***

Prior to the GWA and Whole Exome Sequencing (WES) eras, linkage studies and autozygosity mapping represented the core of genetic analyses. To obtain the first genetic map, restriction fragment length polymorphisms (RFLP) were utilized as landmarks, followed by highly polymorphic microsatellites, typically amounting to 200-400 total genetic markers scattered throughout the genome that were used for mapping traits.<sup>4</sup> For highly penetrant Mendelian diseases, this approach proved extremely valuable, centering on the observation of which genetic markers co-segregated with disease among affected and unaffected family members, thus indicating the genetic region most likely to contain the underlying genetic mutation. Notably, the scientific beauty of genomic linkage studies is elegant: a truly unbiased approach, which can be applied to autosomal dominant, recessive or X-linked modes of inheritance.

Linkage studies made enormous gains upon the completion of the Human Genome Project, commencing in 1990 and ultimately sequencing the human genome in its entirety, at least in draft form, in 2001<sup>5</sup>. Prior to this even in the early 1990's, linkage successes were apparent. For example, the X-linked hypophosphatemic rickets gene, *HYP*, was first identified through a series of multi-locus mapping constructs.

Linkage studies in the 1990's facilitated the concept of homozygosity mapping, in which small consanguineous families are studied to genetically map recessively inherited

disease haplotypes shared by affected individuals but absent in unaffected family members <sup>6,7</sup>.

This method was relatively rapid, because the underlying idea is simple. In consanguineous families with disease, the mutation is likely to be homozygous, and therefore will be surrounded by homozygous genotypes. If the consanguinity is quite recent, then the disease associated genomic region is large, because there have not been many meioses, and therefore little opportunity for the region to break down. By analyzing markers throughout the genome, investigators could identify homozygous regions, often defined as “runs of homozygosity”, characterizing the term autozygosity mapping <sup>8</sup>. Using the same concept as homozygosity mapping, specific haplotypes across the genome may reside only in affected individuals within consanguineous families, allowing one to identify multiple deleterious regions that may contribute to a single polygenic disorder. For instance, data generated from an autozygosity mapping study among highly inbred families manifesting schizophrenia reported that the odds of schizophrenia increase by ~17% for every 1% increase in inbreeding <sup>8</sup>. Likewise, this can also be measured with a LOD score to compare the likelihood of obtaining the test data if the two loci are linked to the likelihood of observing the same data simply by chance alone <sup>6,9</sup>.

The use of traditional linkage panels using a few hundred microsatellite markers was usually followed by a positional cloning project, as once the region of linkage was identified the investigator then had to identify the genes in that region, and determine if any of those genes contained a disease segregating mutation. Both linkage and the subsequent positional cloning experiments were costly, extremely time consuming, and laborious. Furthermore, extensive family pedigrees are required for informative linkage

analysis, which can be difficult to obtain with regard to both physical sample acquisition and accurate history of relatedness and disease. It is perhaps testament to the perceived importance of understanding genetic mutations that cause disease, that so many of these projects were undertaken and completed. There were a number of advances that increased the speed and efficiency of these projects; the first was the human genome project, which meant that an investigator would know (largely) what genes and exons were within their region of interest – thus following linkage they did not have to discover the genes, only the mutations via resequencing, most typically with Sanger based sequencing. Secondly, highly accurate, highly parallel single nucleotide polymorphism genotyping facilitated the rapid execution of linkage analysis.

#### ***1.3.1.2 Next generation sequencing in monogenic diseases***

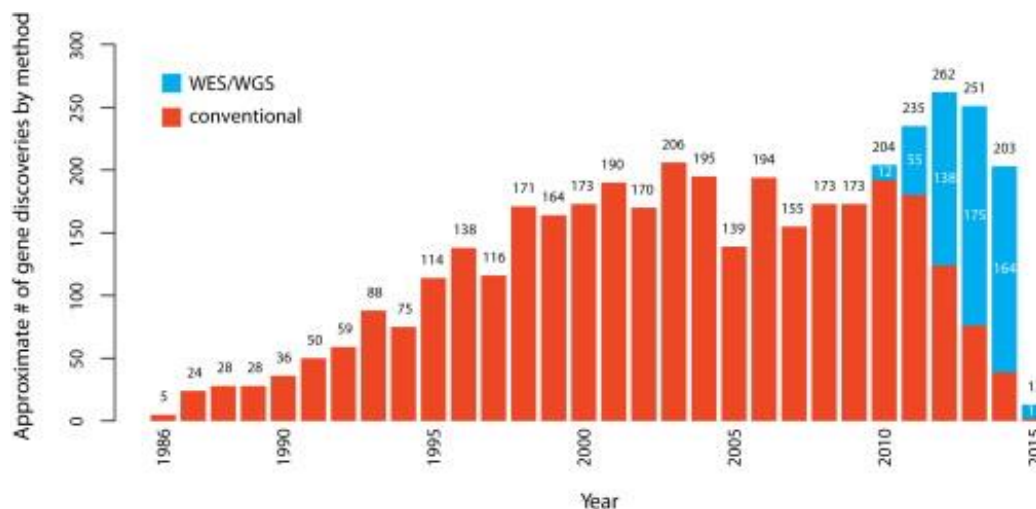
While the function of a large proportion of the human genome remains elusive at present, our knowledge of the central dogma, from DNA transcription to mRNA, and mRNA translation to protein, is largely scientifically sound. Thus, it is logical that we would pursue the variation and functional role of the 1-2% of the human genome containing all coding regions in the form of exons. While this seemingly small proportion consists of approximately 180,000 exons and 27,000 genes, the coding regions are the hot spots of disease causing mutations, as approximately 85% of human monogenic diseases exhibit a causal or associative relationship with missense mutations <sup>10,11</sup>.

The first WES experiment was performed by Hodges et. al in 2007, demonstrating the ability to capture between 55-85% of targeted exonic regions <sup>12</sup>. This provided a significant advance over linkage analysis, which could only demonstrate very large genomic regions co-segregating with disease <sup>7,13</sup> (Figure 1.4). Since then, WES has

proven to be an incredibly powerful approach for not only Mendelian diseases, but additionally for complex diseases particularly in families that are too small for traditional linkage analysis. In consanguineous families harboring runs of homozygosity undetectable via linkage analysis, WES serves as a novel tool to identify these regions, requiring as few as 5 reads of average base coverage (5X) in affected relatives<sup>11</sup>.

Further, WES has revealed novel somatic and *de novo* mutations linked to certain types of cancers as well as early onset PD.<sup>11,14,15</sup> WES can be pursued in families with as few as three individuals (two affected, one unaffected) or even single probands from different families with rare disorders.<sup>16,17</sup>

In addition, WES has also been successfully used to diagnose genetic diseases in individuals lacking known genetic mutations corresponding to his/her presenting phenotype.<sup>11</sup> For instance, WES was first conducted in a single patient who manifested a phenotype consistent with a severe renal salt wasting disease, Bartter syndrome. While candidate genes and variants were identified, none of them had been associated with any known cases of Bartter Syndrome and the patient's diagnosis remained inconclusive. However, upon performing WES of five additional subjects presenting with a similar phenotype, all were shown to carry the same rare deleterious homozygous variant as the proband, facilitating proper diagnosis and pursuit of recommended treatment.<sup>11</sup> Hence, while our knowledge of exome data results is quite limited regarding gene function, the clinical utility cannot be underestimated. The emerging utility of WES and Whole Genome Sequencing (WGS) is evident by the rapid reporting of new genetic insights into clinical disease (Figure 3).



**Figure 3: Approximate number of gene discoveries made by WES and WGS versus conventional approaches since 2010.**

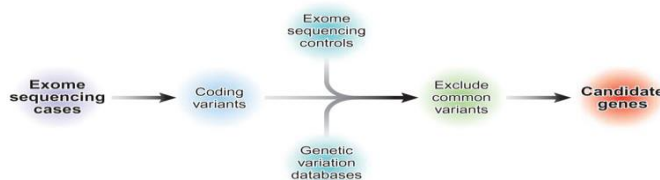
There was an increasing use of conventional methodologies since the mid-1980's until around 2010, whereby WES/WGS protocols became more widely utilized and have since significantly predominated over the use of conventional methods. The first few years of implementing NGS protocol have resulted in an increase in the number of disease gene discoveries.

(Reproduced from Chong et. al. 2010).<sup>18</sup>

To compile such a wealth of information, the exome variant server (EVS) database has been formed: <http://evs.gs.washington.edu/EVS/>. To facilitate the scientific community's understanding of the data, the 1000 genome projects consortium has designed a genomic map to specify the location, allele frequency and local haplotype structure of roughly 15 million SNPs, one million short indels and 20,000 structural variants, most of which were formerly unknown.<sup>19</sup> This has proven to be highly valuable in terms of experiment cost and design, as it has revealed that each person, on average, possesses approximately 250-300 loss of function variants in annotated genes, while 50-100 variants that were previously associated with genetic disorders have been further validated.<sup>19</sup> Further, an enormous database depicting all coding variants sequenced in ~62,000 individuals has been created to help identify the allele frequency and predicted

pathogenicity of many variants not included in 1000 genomes (EXAC

<http://exac.broadinstitute.org/>). A general approach to using WES is depicted in Figure 4.



**Figure 4: Whole exome sequencing analysis schema.**

Whole exome sequencing analysis compares coding variants between cases and controls by using genetic variation databases to filter out common variants, ultimately deriving a list of candidate genes.

Reproduced from (Biesecker 2010).<sup>18</sup>

While there continues to be a need and use for WES, WES does have several limitations. Firstly, while many variants are suspected to be located in coding regions, we have learned that certain diseases harbor variants located within non-coding introns. For example, WES was used to identify unique coding variants as a cause of Kohlschütter-Tonz Syndrome (KTS), a rare autosomal recessive neurodegenerative disease defined by epilepsy, psychomotor regression and amelogenesis imperfecta<sup>20,21</sup>. While a homozygous frameshift deletion and missense mutations in *ROGDI* were determined to be the cause in some affected individuals, coding variants could not explain other cases of disease outside a few families.<sup>20,21</sup> Unsuccessful WES attempts in other families led investigators to pursue WGS, in an effort to seek coverage of regions missed by WES: introns, GC rich repeats, long CNVs and long repetitive sequences. Analysis revealed a homozygous intronic variant, which eliminates the splice donor site within intron 2, rendering a full exonic deletion of exon 2 in *ROGDI* and thus causing disease.<sup>22</sup>



Likewise, while several mutations in coding variants have shown to cause ALS, other forms of alleged Familial ALS (FALS) did not exhibit any known causal variants following WES.<sup>23</sup> However, deep resequencing revealed a large hexanucleotide repeat expansion within the first intron and promoter of *C9ORF72* and has been determined to be a large risk factor for both ALS and Frontotemporal dementia (FTD).<sup>23</sup> Such detection challenges have also occurred for regions enriched with GC content, requiring a combination of cloning, sanger sequencing and *de novo* assembly to reveal the pathogenic variant in medullary cystic kidney disease type 1 (*MCKD1*).<sup>24</sup>

Finally, as mentioned above, CNV detection can be difficult in WES data. Thus, CNVs in Autism were detected by comparative genomic hybridization followed by a series of other assays including microsatellite genotyping.<sup>25</sup> While this does not suggest that WES should not be performed, as it has identified several unique variants associated with autism disorder, it should be used in addition to other techniques to seek a more comprehensive analysis.<sup>26</sup> Hence, one must not assume that negative WES results indicate that casual variants are located in non-coding regions, but rather acknowledge the possibility that such regions were not captured or covered sufficiently.

Finally, psuedogenes can also pose issues when reading short sequence reads obtained from WES. Interestingly, upon learning of *GBA*'s role as an intermediate risk allele for PD, many were interested in detecting *GBA* carriers to enhance our understanding of PD pathophysiology.<sup>7</sup> However, a pseudogene for *GBA* exists only a few kilobases downstream of the target *GBA*, exhibiting approximately 96% sequence homology. Thus, the origin of WES reads at *GBA* are difficult to discern, as they may originate from the pseudogene rather than the intended *GBA* target gene.<sup>7</sup>

Ideally, WGS will become the standard form of next generation sequencing, given the ability of this approach to remedy many of the limitations of WES. The expense of WGS, both in execution and data processing/storage, means that WES is still the dominant methodology; however, history suggests that the price of WGS will drop significantly, and this will shortly be the dominant genetic method.

### **1.3.2 Complex diseases**

With this remarkable progress in the field of genetics, we have also learned that only a small percentage of human diseases are associated with a classical Mendelian inheritance pattern.<sup>18</sup> The majority of disease is believed to be driven by a more complex interaction of factors, with many disorders believed to be a result of an interaction of many genetic variants and environmental factors. These clinical conditions are classified as genetically complex disorders/diseases.

The genetic portion of complex disease is often ascribed to two non-mutually exclusive ideas: The Common Disease Common Variant (CDCV) hypothesis, and Common Disease Rare Variant (CDRV) hypothesis (also called the Multiple Rare Variant (MRV) hypothesis).

In the former, it is postulated that the genetic basis of common complex diseases is a result of a large number of common variants, that each exert relatively small effects on disease risk, but that cumulatively confer significant risk. The risk of these alleles is, individually quite small, because otherwise they would likely have been selected out of the population over time. Notably, the pressure for selection may be relatively low in diseases that occur past reproductive age, or for variants where they may be an earlier selective advantage (balancing selection), thus common risk variants of considerable

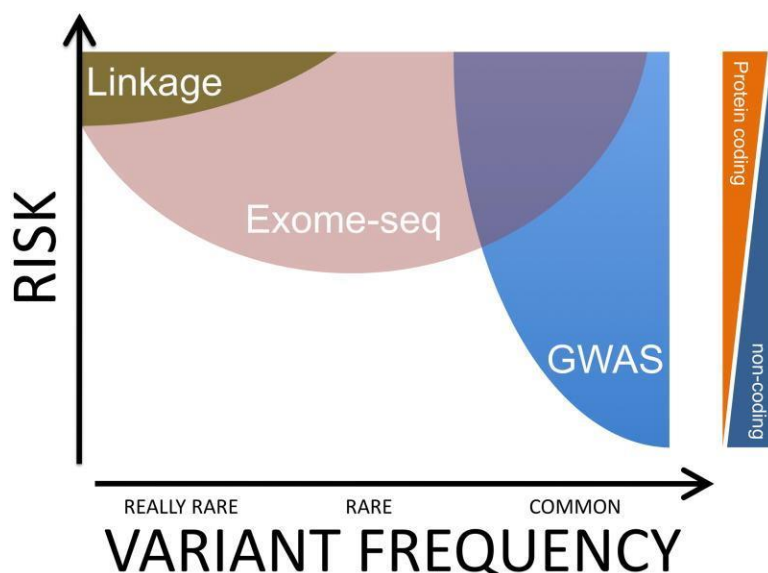
effect do exist (*APOE* in Alzheimer's disease and *CFH* in Age Related Macular Degeneration), however these are the exception rather than the rule. Conversely, the CDRV hypothesis posits that common diseases may be caused by an accumulation of rare variants. Notably, because these variants are by definition rare, they will not have been selected out of the population, even if they have a large influence on disease. While the CDRV and CDCV hypotheses state opposing mechanistic views, they are not mutually exclusive, and it is very likely that both have a role to play in the majority of common diseases.

Finally, to complete the current understanding of variants and graded risk, we must consider the outliers, which seem to defy the current trend of an inverse relationship between MAF and risk with respect to disease. Perhaps the most classic example of this is *APOE* and AD, in which particular alleles of *APOE*, acting in a dose-dependent manner, have been deemed both common and high risk on all GWA studies of AD.<sup>16</sup> This is particularly interesting, as we would expect deleterious common variants to exhibit low reproductive fitness via negative selection over time. However, because AD as well as many other neurodegenerative disorders are late-onset diseases, this logic becomes invalid, as fitness is not affected until much after child-bearing years.<sup>16</sup> The degree of risk is manifested in a dose dependent fashion, in which a homozygous genotype of the highest risk allele, *APOE4*, raises one's likelihood of AD development by approximately 8-fold.<sup>27</sup> While there is clearly other risk in addition to *APOE* isoform carrier status, it is important to study all associated loci to determine an overall graded risk.<sup>16</sup>

The possibility of *de novo* mutations, in which neither parent harbors the variant, also warrants consideration.<sup>15</sup> By the same argument presented above in reference to *APOE*, most *de novo* mutations tend to be highly pathogenic and would ultimately be detrimental towards reproductive fitness. However, our current fund of knowledge regarding *de novo* mutations suggests that their influence commences upon conception; hence, while some mutations may be embryonically lethal, others may render one's fitness very poor from inception and thus can explain early-onset disorders.<sup>15,18</sup>

In referring back to Manolio et al.'s figure (Figure 1), there are also variants that may be very rare and low risk, which would be extremely difficult to detect with current technologies, but we must acknowledge the possibility of their presence. While we would expect them to play a role in the graded risk equation, their low risk profile and challenging detection status hinder their center stage presence in our current analysis of the genetic landscape underlying human disease.<sup>1</sup>

According to variant frequency and hypothesized degree of risk harbored by the gene of interest, one must utilize an appropriate method of gene discovery (Figure 5).



**Figure 5: Gene discovery methods.**

Methods based on the hypothesized nature of the genetic architecture of the disease under investigation. Very rare protein coding, high-risk variants are best detected using either linkage analyses or WES. However, WES is likewise able to identify common variants, whereas linkage use is amenable only to extremely rare variants ( $MAF < 0.001$ ). GWA studies are able to detect both coding and non-coding common variants with varying levels of risk.

Reproduced from (Singleton, et al., 2010).<sup>2</sup>

### ***1.3.2.1 Candidate gene association studies in complex diseases***

When penetrance is incomplete, certain approaches (i.e. linkage analysis) can be particularly ineffective, as unaffected family members may possess the genotype but not express the corresponding phenotype.

A candidate gene approach can be implemented when there is prior knowledge of gene function and a possible role in disease. A list of putative candidate genes is formed by including those that are both functionally relevant and located at plausible genomic loci (i.e. signal detection from previous GWA study). Further, using online tools, such as KEGG and Ingenuity, which provide graphical pathway maps, or STRING, which reveals all known and possible gene interactions, one can generate a substantial candidate gene list. As there is extensive overlap in genes causing complex diseases, many of these

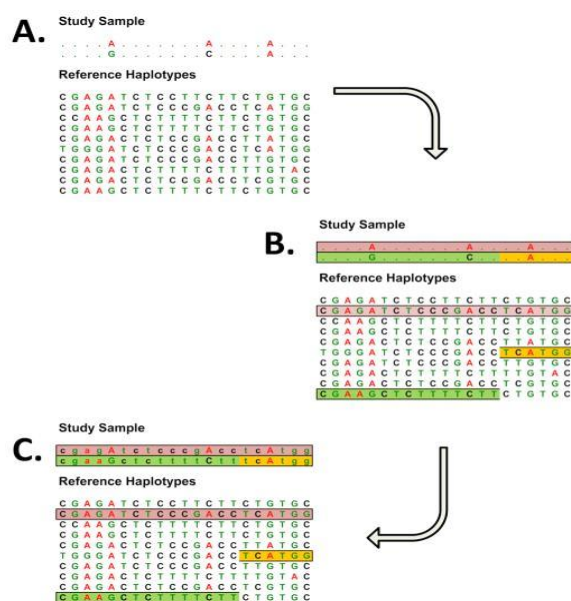
associated genes often become candidates for diseases in the same family (i.e. *SNCA* in alpha-synucleinopathies: Parkinson's Disease and Multiple System Atrophy). An important caveat, nonetheless, is that candidate gene analyses require *a priori* hypothesis of the genes selected for investigation and a limitation for such efforts is a lack of understanding of the underlying biology of disease, of the likely size of effect of risk variants, or the location of variants. Given that we often know little about the disease process, and that many of the initial candidate gene association studies were small in scope and power, it is not surprising that there has been an abundance of false positive (type I error) reports in which initially published associations were unable to withstand independent replication.<sup>7,28</sup> Likewise, underpowered studies may fail to reveal significant loci, such as the case for initial studies for PD, which reported conflicting results for both *MAPT* and *SNCA*, now known to be risk loci. Hence we must also entertain the possibility of false negatives (type II error).

### ***1.3.2.2 Genome wide association studies in complex diseases***

The primary pitfall of candidate gene association studies were that they were inherently biased, and generally rather small, perhaps in large part because at the time most investigators were looking for effects of the size associated with *APOE* in AD. The answer to identifying common risk variants for disease lay not in focused candidate studies, but in an unbiased method (much as linkage is unbiased). Two advances allowed progress to be made in this regard; the development of highly parallel and accurate SNP genotyping arrays, and the International Haplotype Map Project.

In 2002, the HapMap 1000 genomes International consortium was established to study the common patterns of DNA sequence variation across the human genome. This

compilation entailed analysis of sequence variants, their frequencies and respective correlations between them among populations from several continents, Europe, Asia and Africa (International hap map consortium 2003). While the Human genome project focused on sequencing the entire human genome, including the 99.9% of DNA which all humans share, the HapMap project targeted the distinct variation among diverse population cohorts in the remaining 0.1% <sup>29</sup> This facilitated the process of imputation, which uses a reference panel of whole genome sequenced samples to estimate genotypes in positions that were not included among the markers in the genotyping assay (Figure 6).



**Figure 6: A schematic of how imputation works**

A.) Observed genotypic data from unrelated individuals is incorporated into a HapMap reference panel (B) which detects shared chromosomal regions between study samples and those in HapMap reference panel. For samples of European ancestry, haplotype stretches are typically >100kb in length. C.) Haplotype sharing information is combined with observed genotypes of samples to “fill in” unobserved genotypes in study samples.

Reproduced from (Li et al 2009).<sup>30</sup>

These advances drove the new era and concept of GWA studies, which are based on the following premise: risk variants may occur within haplotype blocks shared with common variants through linkage disequilibrium (LD). A key concept in population genetics, LD refers to the association of alleles at distinct loci in a non-random fashion; thus, alleles in LD are associated much more often than would be expected by chance (linkage equilibrium). Since common variants can be tagged through genotyping marker arrays, risk variants in LD should manifest an association, by proxy, with tagged common variants and ultimately with the disease trait under investigation.

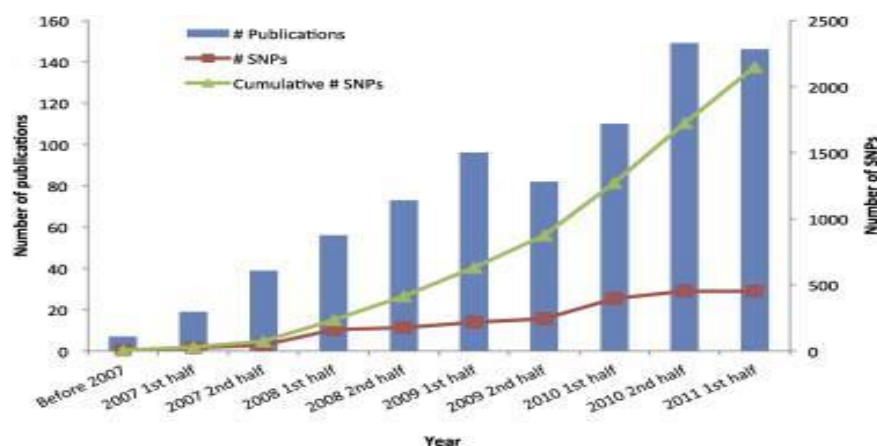
While critics suggest a metaphor of a genome wide fishing expedition,<sup>2</sup> the overwhelming majority of the genetics community would argue that the results gleaned from GWA studies mark a significant advancement from linkage based analyses.<sup>1,2</sup> Specifically, many believe that GWA studies have modified our approach toward experimental design while demonstrating higher power than linkage studies to detect common variants with mild effects.<sup>7,9</sup> By using such an unbiased approach, previous knowledge of genomic structure and trait etiology are not considered *priori* hypotheses, which has ultimately yielded key information regarding common disease pathways for several disorders.<sup>7</sup>

It is evident that understanding population heterogeneity plays a crucial role in the ability to successfully implement a GWA study, as varying genotypes among ethnically distinct populations could easily be interpreted as false positive associations with the disease trait under investigation. Moreover, the HapMap project has also revealed extensive levels of intra-population heterogeneity in those populations characterized by a history of mass migration and minimal isolation.<sup>31,32</sup>



Notably, the first successful GWA study only assessed 100,000 SNPs in 96 patients affected with macular degeneration and 50 controls (Klein et al 2005). Despite its small size, SNPs in the complement factor H (*CFH*) gene were deemed highly significant and later confirmed with replication.<sup>33</sup> Since its earliest success, GWA studies have been performed in more than 200 diseases.<sup>34</sup> Further, the concept of GWA studies has been applied to scrutinize variants which contribute to normal variability in human traits (i.e. height).<sup>35</sup> The vast amount of data generated from all GWA studies to date are organized in the publically accessible online catalogue (<http://www.genome.gov/26525384>).

Visser et al has depicted the number of GWA studies that have been performed since 2005, illustrating the continuous increase in the cumulative number of SNPs incorporated with each additional investigation (Figure 7).



**Figure 7: Number of GWA studies published per year.**

The cumulative number of SNPs included in GWA studies since 2007 has increased dramatically. The corresponding number of SNPs and publications have likewise continued to increase but in a more linear fashion. Single nucleotide polymorphisms (SNPs) with p-values <10<sup>-8</sup> are illustrated.

Reproduced from (Visser, et al., 2012).<sup>35</sup>

The underlying premise for pursuing GWA studies is based on the CDCV paradigm, with the goal of detecting common variants ( $MAF > 5\%$ ) that contribute to the development of disease.<sup>1</sup> By using commercialized SNP chips or arrays that capture the majority of common variation throughout the human genome in studies of at least 1000 cases and controls, a minimum of 300,000 markers has been suggested to obtain statistically significant results, although clearly these parameters vary extremely between diseases.<sup>36</sup> The significance of the *CFH* gene in the macular degeneration study is a clear exception, as the signal intensity was highly detectable despite low power.<sup>33</sup> Likewise, this unique combination of common frequency with a high graded risk profile also applies to *APOE* detection in AD.<sup>27</sup> PD GWA studies, however, are an excellent example of the concept of sheer power; using numbers below the ideal sample and SNP marker levels, loci which we know are definitive PD risk factors (*PARK16*) did not reach significance. However, by simply increasing sample size and genotype marker frequency, lower grade risk variants reached statistical significance in the PD GWA study meta-analysis, thus overcoming the initial low power issue in smaller studies.<sup>37</sup>

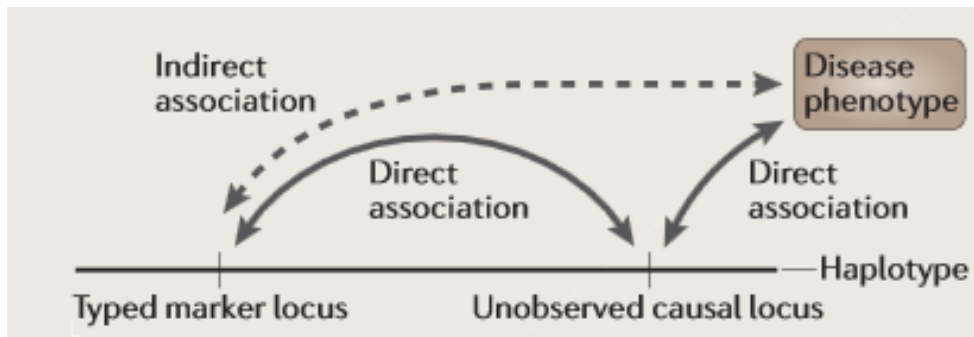
Greater than 80% of known associated variants lie outside of coding regions, which highlights the importance of surveying both coding and non-coding regions for plausible disease-associated variants.<sup>1</sup> While this is beneficial towards a more comprehensive genome wide analysis, the fact that many of the non-coding regions have poorly (if any) defined functional knowledge can be challenging. To remedy this issue, the Encyclopedia of DNA Elements (ENCODE) project was developed in 2007 to analyze the 1% of functional elements in the genome, which are often non-coding but may effect transcriptional activity or regulate splicing.<sup>38</sup>

While GWA studies are targeted for common diseases based on the CDCV paradigm, some common diseases harbor variants with such a mild graded risk that the sample size required for detection is simply not feasible.<sup>2,39</sup> In addition, many complex diseases exhibit allelic heterogeneity, whereby distinct disease causing mutations exist within a single gene (i.e. *LRK2*). While some may be highly penetrant (p.*G2019S*), others may require multiple alleles to co-exist in LD and be inherited as a haplotype block to consequently cause disease.<sup>40</sup> Thus, while we have discovered several common variants for diseases like PD, there are likely others we have yet to identify.

In addition to increasing sample size, increasing the number of genetic markers may also improve the odds of successfully identifying association. A critical step forward in this was the development of imputation, which is used to increase the power of GWA studies. This is not only beneficial towards enhancing the power of signal detection, it is often a key step in order to strengthen fine mapping abilities and mitigate the effects of possible synthetic associations.<sup>7,16</sup>

Without imputation, power is often insufficient to distinguish signals deriving from single markers due to LD between variants that are physically near one another.<sup>16</sup> These may be considered synthetic associations, which refer to the indirect associations that can occur between a common variant and at least one or more rare causal variants.<sup>41</sup> Hence, a positive signal derived from a GWA study is frequently not located within the functional domain of any gene. Consequently, the most nearby and biologically plausible gene is often declared the candidate gene, but this must not be deemed associative until independent replication and functional work are performed to confirm its physiological role.<sup>16</sup> In a Sickle Cell Anemia GWA study among the Yoruba Ibadan Nigerian

population, a total of 189 SNPs were deemed significant, encompassing a 2.5Mb region on chromosome 11.<sup>41</sup> The strongest association signal was 9kb from the closest gene, *OR51V1*, which is not even the causal gene of Sickle Cell Anemia. While *OR51V1* is very close to the causal gene, *HBB*, this demonstrates how strong signals can migrate across several LD blocks towards distant areas of the genome.<sup>41</sup> This concept is illustrated in Figure 8:



**Figure 8: Direct and indirect nature of associated variants.**

When a typed marker locus and an unobserved causal locus are in the same haplotype block, they can both appear to demonstrate a direct association with the disease phenotype, by proxy. However, the typed marker locus truly exhibits an indirect association with the disease phenotype due to LD, thus creating a synthetic association.

Reproduced from (Balding et al, 2006).<sup>36</sup>

If a candidate locus can be identified by a GWA study signal, it is recommended to use deep resequencing within a +/- 10Mb region surrounding the locus to determine the signal origin.<sup>41,42</sup>

While false positives are a valid concern, the probability of false negatives becomes increasingly likely with the performance of each additional statistical correction test for multiple testing.<sup>43</sup> The Bonferroni correction method, which is the most widely used and accepted correction method to determine GWA signal significance, is often faulted for being overly conservative due to the following: the alpha ( $\alpha$ ) value denoting

the signal significance is divided by the number of independent tests performed, which does not account for LD among adjacent SNPs or even genes associated with a known, common biological pathway.<sup>43</sup> As a result, this increases the type II error rate (false negatives) and reduces the specificity for signal detection.<sup>43</sup> Furthermore, it is important to recognize that many of the rare variants not captured by GWA study arrays are likely not included post imputation. While there are notable exceptions, including the detection of rare variants in lung cancer after imputation, this is an unlikely scenario.<sup>39</sup> Hence, when imputation is insufficient to attain desired power or does not yield significant signals, it is advised to seek replication through direct genotyping in an independent cohort.<sup>44</sup>

Secondly, even if sample sizes are adequate, some complex common diseases are extremely heterogeneous in nature; hence, it is quite possible that risk variants are specific to only certain subtypes of disease. Thus, the translation of this information from clinical diagnosis to a classifiable phenotype for research purposes is often hazy at best. Consequently, the current clinical utility of GWA study results has been far from optimal.<sup>45,46</sup>

### ***1.1.2.3 Estimating Risk***

Classically, two methods have been used to estimate heritability: twins studies and relative risk. Regardless of approach, heritability is defined as the proportion of total variance in a population for a specific measurement, obtained at a specific age or time, which is attributable to variation in additive genetic or total genetic values.<sup>47</sup> Measurement is obviously quite challenging and always an approximation, as it is

dependent upon segregation of alleles affecting the trait in the population, frequency of alleles, effect sizes of specific variants and the mechanism of action underlying particular genes (i.e. epistasis). However, by comparing the observed and expected concordance of particular binary traits among relatives, we can quantify estimates of heritability. The closer the genetic relationship between individuals, the less genetic variance is expected, allowing one to better characterize the role of non-genetic influences (i.e. environmental, epigenetic). However, even in monozygotic twins with a very high concordance rate for a particular disease not confounded by shared environments, a high heritability still does not reveal information about the genetic architecture of the traits or even how many loci contribute to the phenotype of interest.<sup>47</sup>

When assessing the likelihood of disease risk in medicine, the relative risk is calculated, which is another form of measuring heritability. By comparing the probability of an event happening (i.e. development of disease) in distinct individuals or populations, we can learn who may be at a higher risk. While increased risk may be attributed to environmental or behavioral/lifestyle factors (i.e. smoking), it also may be due to additive genetic factors shared by individuals harboring similar risks. In the latter, we can estimate heritability from these individual genetic loci to better understand persons at risk. This illustrates the calculation of the odds ratio in GWAs studies by determining risk alleles and the corresponding odds of developing disease in particular individuals or populations based on the presence or absence of specific genetic variants.

In addition to detecting risk loci, GWA study data have been used to assess heritability levels of complex diseases through polygenic additive inheritance via Genome Wide Complex Trait Analysis (GCTA).<sup>48</sup> This allows ones to compare the

genotypes of all common variants (included on the SNP array) between cases and controls in a particular disease of interest in an effort to explain phenotypic variance attributed to genotypic differences.<sup>48</sup> Notably, this has been performed for several complex diseases and there is a large discrepancy between the estimated heritability derived from GCTA results, which does not account for Single Nucleotide Polymorphism (SNP) effect size, with the accumulation of signals detected from the corresponding GWA study. While this is initially perplexing, this concept of missing heritability can be explained through a variety of reasons.

Firstly, the inability of GWA to detect rare variants (MAF<5%) (with low or high degrees of risk) would explain the low levels of heritability gleaned from GWA identified loci on their own.<sup>1,16</sup> Additional explanations include but are not limited to the inability of genotyping arrays to capture structural variation in samples including copy number variants (indels), copy neutral variants (inversions and translocations) and repeat regions, as well as epistasis.<sup>1</sup> For example, while autoimmune diseases like Crohn's and Psoriasis have demonstrated an association with common CNVs that harbor modest effects, neuropsychiatric conditions like Autism and Schizophrenia are associated with rare CNVs that exhibit large effects.<sup>1</sup> Hence, using a simplified genetic model to estimate heritability that fails to account for SNP effect size and LD, as well as environmental factors (which remain largely elusive), estimated heritability results must be interpreted with great caution.<sup>1,49</sup>

The knowledge of disease variants and their plausible biological function derived from linkage analyses and GWA studies continues to grow rapidly. Key theories have been confirmed towards our evolving comprehension of genetic disease: non-causal rare

variants imparting moderate or high risk obtained through linkage studies in support of CDRV, and common variants portraying mild, moderate or even high risk (i.e. AD and APOE) derived from GWA studies endorsing CDCV are both necessary but not sufficient to paint the full genetic portrait of human disease.<sup>1,2</sup> Given that our current heritability estimates, which we acknowledge must be interpreted with great caution, challenge our current domain of knowledge derived from GWA and linkage studies, the road to discovery is only in its infancy. Thus, we have continued to forge ahead into our newest and most exciting genomic technology to date: whole exome and whole genome sequencing.

### ***1.3.2.3 Next generation sequencing in complex diseases***

WES has not only helped identify and properly diagnose monogenic diseases, as in an atypical case of Wolfram syndrome, as well as Freeman-Sheldon syndrome<sup>50–52</sup>, but holds promise for discovering both rare and common variants among patients with known polygenic disorders. While diseases with classic Mendelian forms of inheritance serve as ideal candidates for exome sequencing, one must realize that complex disorders, provided that sample sizes are sufficient, are potentially amenable to dissection with WES.

The identification of *TREM2* variants as risk factors for Alzheimer's disease is an excellent example of this approach.<sup>44</sup> Investigation of potentially causative genetic loci of complex diseases requires acknowledging the concept of variable expressivity. In essence, variable expressivity is considered the “rule” rather than “exception”; hence, phenotypic variation may result in discordance among genotype-phenotype assessments even among highly penetrant mutations.<sup>53</sup> Analysis of GWA study results for common



diseases suggests that the majority of heritability behind complex traits is unlikely to be attributed to common variants with mild effects alone but rather, that a significant proportion of the heritability associated with complex diseases is likely to be attributed to rare variants, which as discussed above, may also have larger effect sizes.<sup>54,55</sup>

While WES can be performed on a variety of studies, the filtering process must be tailored accordingly. In the case of trios, in which both parents are unaffected and the child is affected, a homozygous recessive or *de novo* mutation would be expected in the child. Upon data generation, one would focus on variants harbored by the child (in homozygous form) while each parent is heterozygous. If this does not yield promising results, the hunt for a *de novo* mutation, whereby the child is heterozygous for a novel allele and both parents are homozygous wildtype, is an alternative filtering strategy. Furthermore, some WES analyses have been able to obtain multiple sets of trios with the goal of identifying a rare and novel variant shared by affected children in different families. This particular strategy has been very successful in the identification of several novel variants causing familial ALS, including Valosin-containing protein (VCP) and Matrin 3 (MATR3).<sup>56,57</sup>

Other studies involve the analysis of multigenerational families, often with an unknown pattern of inheritance. However, based on the presence or absence of disease “skipping generations,” or predominance of one sex manifesting disease, one can hypothesize possible modes of inheritance and filter accordingly. Using two affected yet distantly related individuals within familial pedigrees can greatly decrease the number of candidate genes and loci from WES results. Such a strategy has been key in the

identification of a rare variant within vacuolar protein sorting 35 (*VPS35*) as a cause of familial PD.<sup>58,59</sup>

While obtaining families is ideal for genetic analyses, they are often hard to obtain with accurate clinical and relatedness histories. Therefore, many WES analyses acquire numerous sporadic cases and perform a case control analysis based on age and population matched controls. While this approach has also been successful, such as the *TREM2* discovery in AD, the potential for heterogeneity between affected samples is significantly higher and further complicates genetic analyses in comparison to strictly using familial cohorts.<sup>60</sup>

In comparison to GWA studies, which measure statistically significant associations using an odds ratio (OR), WES filtering for very rare variants is assessed through minor allele frequency (MAF). MAF exhibits an inverse linear relationship with a required sample size, in which  $1/\text{MAF}$  is directly proportional to the sample size. Thus, it is evident that substantially large cohorts are the most promising towards finding such rare variants.<sup>1</sup> However, sample size demonstrates a quadratic relationship,  $1/|(\text{OR}-1)|$  with the odds ratio, which is necessary for association detections. Therefore, prior association studies (measured by OR) have all needed a significantly larger sample size than WES (detection measured by MAF) since sample size is much more strongly affected by OR than MAF.<sup>1</sup> Thus, even when cohort numbers are in the hundreds (vs. thousands), WES is an invaluable tool and is therefore more likely to be lucrative than an association study with the same sample size.

While sample acquisition is challenging when studying any rare disease, particularly those that require pathological confirmation, the ability to utilize fewer

samples within WES (vs. GWA) studies provides increased power and opportunity to reveal statistically significant associations through individual variant and gene burden analyses. In a classic case control study, WES and WGS allow one to uncover both protective and deleterious alleles (and possibly genes) in the pathogenesis of disease. Upon identification of putative associated or causal variants, one must validate through traditional sequencing methods (i.e. sanger sequencing) and replicate these results in independent cohorts.

As with any new technology, there are some limitations of WES that must be addressed: firstly, incomplete current capture efficiency means that the remaining exomic regions are not captured nor sequenced. Secondly, given that WES only targets coding regions, it cannot detect intronic regions involved in gene regulation or expression. Thirdly, given its bias to coding regions only, WES cannot characterize all genomic structural variation.<sup>61</sup> Finally, since WES is a research tool in its infancy and likely to be clinically unavailable for many years, financial investment towards reagents and equipment cannot be overlooked. Notably, however, costs have declined exponentially in recent years.

While attempts to confront some of these issues are simply not feasible due to technological constraints (i.e. capture inefficiency and inaccuracy of variant calling), other issues, such as insufficient sample size, can be addressed. Perhaps the most obvious solution to this problem requires international collaboration and data sharing. With the ability to exchange petabytes of data through universal drives such as amazon cloud, our current progress in the field of genetics and specifically in overcoming certain limitations of WES is profound. Thus, in an effort to efficiently and easily share resources, the

burden of analysis, and rapidly disseminate results, the formation of an international collaborative framework should be the priority of any entity wishing to pursue research into the genetic basis of very rare disorders like MSA.

### **1.3.3 The future of human disease genetics**

Many genetic variants that underlie disease have been identified for which a pathobiological function is unknown; thus it has been argued that time and resources would be better spent understanding the biological basis of these factors rather than identifying more. A converse opinion is that accumulating additional genetic risk for disease provides additional understanding of the disease process as a whole, and thus the way to understand the mechanisms of disease is to identify as much of the genetic influence for this disease as possible. With this argument in mind, much of the work in this thesis centers on attempts to further understand the genetic basis of two devastating neurodegenerative diseases PD and MSA. The pursuit of genetic risk and causative loci is scientifically tractable via the implementation of next generation methods, including GWA and second generation sequencing, provide the ability to obtain valuable data for disease investigation and to inform clinical diagnosis.<sup>53</sup>

Success in any modern genetic investigation requires extensive scientific collaboration, regardless of approach. In particular, diseases such as MSA are rare enough that no single group can collect sufficient cases on its own; thus, the field will not progress without pooling of clinical resources. Upon acknowledging these substantial challenges discussed above, we would predict that clinical progress of PD and MSA (diagnosis, treatment) will be much delayed until we make advances towards our genetic understanding of these diseases.

## 1.4 Parkinson's disease

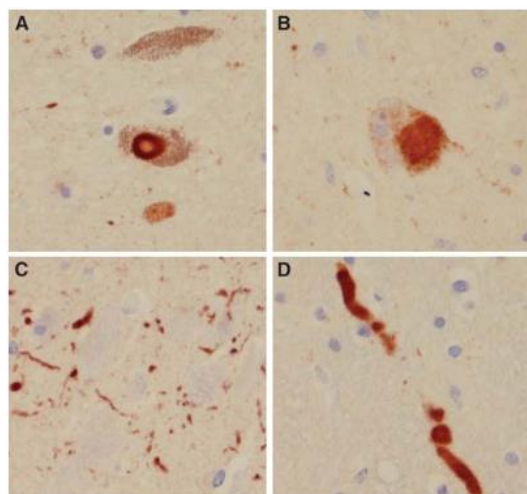
Second to AD, PD is the most common neurodegenerative disorder, with an approximate prevalence of 50-200 cases for every 100,000 individuals worldwide.<sup>62</sup> PD demonstrates an age-dependent prevalence in which roughly 1% of the global population is affected by 65 years of age, while approximately 4-5% of individuals at 85 years of age suffer with PD.<sup>63</sup> The average age of onset is variable but is approximately 70 years of age. Up to 10% of PD patients develop disease prior to 50 years of age, representing those individuals with familial forms of disease.<sup>62</sup> PD has been characterized as a complex polygenic disorder that is influenced by both genetic and environmental factors.<sup>62,64</sup>

The use of levodopa remains the universal first line of treatment for PD. While levodopa, typically improves a patient's parkinsonian symptoms, this period is usually temporary and increased doses are required to maintain efficacy. Ultimately, even if levodopa helps alleviate symptoms, it does not stop or even slow down disease progression, making the need for targeted disease modifying therapies in PD a priority.<sup>65</sup>

### 1.4.1 Clinical and neuropathological features of Parkinson's disease

Classical PD refers to a patient presenting with four key clinical features: bradykinesia, resting tremor, muscle rigidity and postural instability.<sup>65</sup> While these neurological symptoms define typical PD, patients often present with an array of non-motor disturbances including (but not limited to) sleeping maladies, constipation, depression, progressive dementia and orthostatic hypotension.<sup>66</sup> Autopsy is required for definitive diagnosis of PD, which must reveal PD's pathognomonic hallmark: Lewy body

(LB) inclusions. While there is some heterogeneity depending on the distinct etiology of PD, postmortem PD brains typically reveal significant neuronal death in the (SNPC) nigra pars compacta with alpha-synuclein filled LB inclusions permeating surviving neurons.<sup>65</sup> In addition to the formation of LBs, alpha-synuclein also accumulates in neuronal processes, called Lewy neurites (LNs).<sup>67</sup> (Figure 9, AB)



**Figure 9: Microscopic findings in PD.**

Microscopic findings in PD with alpha-synuclein immunohistochemistry. A typical brainstem type Lewy body (A) and a pale staining “cortical type” Lewy body (B), Lewy neurites in CA2 sector of hippocampus (C), and intraneuritic Lewy bodies in medulla (D).

(Reproduced from Dickson 2012.)<sup>67</sup>

PD neuropathology also extends outside of the brainstem, often noted in the hippocampal and medullary regions (Figure 9) When LBs are identified within cortical regions of the brain, such as the amygdala, they are known as cortical LBs.<sup>67,68</sup> Furthermore, pale bodies, which are defined as pale staining neuronal cytoplasmic inclusions localized to the brainstem, are another key feature of PD neuropathology.<sup>67</sup>

While reduced pigmentation in both the dopaminergic neurons in the substantia nigra and noradrenergic neurons in the locus ceruleus are key features to individuals with PD, specific regions of the brain are typically affected with respect to the underlying genetic mutation<sup>69</sup> (Table 1).

Locus/gene	Mutation	Neuropathology
<i>PARK1,4/SNCA</i>	p.A53T, p.A30P, p.E46 K	Nigral neuronal loss, cortical and brainstem Lewy bodies
	SNCA triplication/duplication	Cortical and brainstem Lewy bodies, temporal lobe vacuolation.
<i>PARK2/PARKIN</i>	Exon4Del/Exon4Del	Nigral neuronal loss, no Lewy bodies
	p.K211 N/Exon4Del	
	p.Q34fs/p.Q34fs	
	Exon3Del/Exon3Del	Nigral neuronal loss, no Lewy bodies, $\alpha$ -synuclein positive inclusions in the pedunculopontine nucleus
	p.R275W/p.Pro113fsX51	Cortical Lewy bodies, none in the brainstem but occasional Lewy neurites in the dorsal nucleus of vagus.
	Del1072T/Exon7Del	Lewy bodies in the locus ceruleus and substantia nigra
<i>PARK8/LRRK2</i>	p.G2019S	Neuropathology ranges from non-specific nigral degeneration to widespread Lewy body disease.
	p.R1441C	Varies from Lewy body disease to nigral degeneration with ubiquitin positive inclusions to severe tau pathology
	p.R1441G/p.I2020T	Non-specific nigral degeneration
	p.Y1699C	Varies from Lewy body disease to nigral degeneration with ubiquitin positive inclusions or Alzheimer pathology
<i>PARK6/PINK1</i>	Exon7Del/c.1488 + 1G > A	Nigral neuronal loss, Lewy bodies and aberrant neurites in the reticular nuclei of the brainstem, substantia nigra pars compacta and Meynert nucleus.
<i>PARK14/PLA2G6</i>	p.R37X/c.1078-3C>A	Range from mild to severe Lewy body disease, with neurofibrillary tangles and axonal spheroids
	p.T572I/p.T572I	
<i>PANK2</i>	p.G521R/p.G521R	No Lewy bodies, diffuse tau pathology

**Table 1: Neuropathology of monogenic forms of PD.**

Neuropathology is grouped according to mutation.

(Reproduced from Houlden et al 2012).<sup>69</sup>

Prior to involvement of the central nervous system, some have suggested that PD initially affects the autonomic neurons within the peripheral nervous system.<sup>67,70</sup> In a case control study investigating the epicardium, researchers observed a significant reduction in cardiac sympathetic denervation in PD cases as compared to controls. Interestingly, using alpha-synuclein immunohistochemistry, a correlation was seen between the density of protein aggregates and disease duration.<sup>71</sup> Further, PD pathology has been observed in the enteric nervous system and submandibular glands, suggesting other plausible avenues of

synaptic cell transmission between the enteric nervous system and central nervous system.<sup>72</sup>

While clinical and pathological features of typical PD are critical for proper diagnosis, the heterogeneous nature of this disorder cannot be underestimated, as LB pathology is neither sufficient nor necessary for clinical diagnosis of PD. It is also notable that patients with the same disease causing mutation may present with and without LB inclusions. On the contrary, the presence of LB pathology is a common feature in several of the atypical parkinsonisms, also known as Parkinson-plus (PP) syndromes.

The clinical diagnosis of such disorders, including MSA, Dementia with Lewy Body disease (DLB), Progressive Supranuclear Palsy (PSP), Corticobasal degeneration (CBD) and Juvenile-onset Pallidopyramidal syndromes is contingent upon the presence of parkinsonian motor dysfunction concomitant with atypical symptoms deemed as PD exclusion criteria. These may consist of hallucinations, dysautonomia, ataxia, dystonia, early dementia and several others.<sup>65</sup> Such atypical parkinsonisms will be discussed in greater detail in the MSA genetic etiology review section in the next section of this chapter.

#### **1.4.2 Genetic etiology of monogenic PD**

PD causing mutations have been identified in 15 genes responsible for Mendelian forms of PD. While up to 10% of cases of PD are familial, many of these exhibit variable penetrance, whereby external factors modify phenotypic presentation.<sup>62</sup> While substantial evidence suggests PD is a multifactorial disease with variable penetrance, we have gleaned key insights from the rare Mendelian exceptions. Prior to the identification of



mutations or CNVs in *SNCA* as disease causal, much of the scientific community was skeptical regarding an association between familial and sporadic cases of PD. However, the discovery of *SNCA* mutations and identification of the protein product of *SNCA*, alpha-synuclein, localized within LBs mitigated critics.

### ***1.4.2.1 Autosomal dominant PD***

#### ***1.4.2.1.1 LRRK2***

As the cause of the most common form of familial PD, mutations in Leucine rich repeat kinase 2 (*LRRK2*), account for the majority of all known heritable PD mutations.<sup>62</sup> Initially identified via linkage analysis, *LRRK2* carriers are usually affected in their sixth decade of life, and manifest clinical features similar to those with sporadic PD, albeit slower progression despite increased frequency of tremor and dystonia.<sup>62,65,73</sup> Although close to 80 *LRRK2* gene variants have been identified among diverse global populations, a mere seven of these have been unequivocally determined to cause disease.<sup>74</sup>

The most common mutation in *LRRK2*, p.G2019S, is responsible for between 5-40% of dominantly inherited or sporadic PD, dependent upon the population under scrutiny. Interestingly, prevalence of this variant exhibits a distinct south to north gradient, in which North African Arab and Jewish populations possess the highest frequency, decreasing as one progresses further north across populations of European descent.<sup>62</sup> Specifically, an extensive case control study reported that p.G2019S demonstrates a prevalence of 0.71%, .07% and 30.25% among Caucasian, Asian and Arabic PD patients, respectively.<sup>75</sup> It is believed that this variant was derived from a common founder between 4500-9100 years ago in the Near East and subsequently migrated globally with the Ashkenazi Diaspora.<sup>76,77</sup> Further, *LRRK2* p.G2019S exhibits a

pattern of age dependence and incomplete penetrance, rising to approximately 75% at 80 years of age.<sup>62</sup>

Resonating with classic PD neuropathology, *LRRK2* mutation carriers typically exhibit neuronal loss in the SNPC and LB inclusions among surviving neurons. Notably, however, the very first case among the Sagamihara kindred in Japan, identified to harbor the p.I2020T mutation, did not exhibit LB pathology.<sup>78</sup> However, additional pathological features may be present such as concomitant neurofibrillary plaque and tangles, anterior horn cell pathology or SNPC neuronal loss in the absence of LB. Moreover, glial cytoplasmic inclusions (GCIs), the pathological hallmark of MSA, have also been reported in PD patient brains.<sup>78–81</sup> Such pleomorphic pathology in the brains of *LRRK2* mutation carriers has been reported within single *LRRK2* families.<sup>79,82,83</sup>

Interestingly, *LRRK2* mutation carriers have been suggested to be at an elevated risk of several types of cancers (i.e. malignant melanoma), while individuals harboring common variants in *LRRK2* demonstrate an association with autoimmune disorders (i.e. Irritable Bowel Syndrome, Crohn's Disease) and leprosy. On the contrary, such individuals have also demonstrated a decreased risk of non-skin cancers (i.e. lung cancer, prostate cancer), further complicating our understanding of *LRRK2*'s complex role in human physiology.<sup>84–88</sup>

*LRRK2* codes for a ubiquitous, multi-domain protein.<sup>62,89</sup> Numerous studies have reported two distinct enzymatic subunits: a kinase domain and a GTPase, interconnected by a COR segment, as the site of most pathogenic mutations.<sup>62</sup> Mutations with the COR segment or enzymatic subunits have revealed *LRRK2*'s role in neuronal growth, cytoskeleton maintenance, vesicle trafficking and chaperone-mediated autophagy.<sup>64,91–93</sup>

#### 1.4.2.1.2 SNCA

Identified through linkage analysis, *SNCA* mutations in the form of duplications, triplications and point mutations represent the second most common cause of autosomal dominant PD.<sup>62,93,94</sup> While the prevalence of disease causal *SNCA* mutations is significantly less than those of *LRRK2*, variation in the *SNCA* locus is a key risk factor for idiopathic PD and vital to unraveling PD pathophysiology.<sup>95</sup> Notably the first mutation identified in PD was in *SNCA* and the subsequent identification of the protein product as a major component of Lewy Bodies elegantly tied together rare genetic and common forms of PD.

Clinically, patients harboring *SNCA* mutations display parkinsonian features in addition to more atypical symptoms such as myoclonus, severe dysautonomia, dementia, and possibly progressive loss of levodopa responsiveness.<sup>62</sup> The moderate prevalence of dementia suggests that PD-dementia and DLB exist on a clinical-genetic continuum.<sup>96</sup> Further, it has also been observed that the disease onset and co-morbidity of severe dementia and psychiatric issues may be associated with the distinct number of *SNCA* copies carried by duplication or triplication carriers. For example, those carrying *SNCA* duplications develop disease around a decade later than those individuals harboring *SNCA* triplications, with the latter characterized by higher levels of severe dementia.<sup>69</sup>

The neuropathology of *SNCA* mutation carriers reveals classic SNPC neuronal loss with widespread LB inclusions in both the brainstem and cerebral cortex. Further, brains harboring pathological *SNCA* mutations may also present with temporal lobe vacuolation.<sup>69,96–98</sup> Given that *SNCA* encodes the protein alpha-synuclein, which has been determined to be a substantial component of LB inclusions, *SNCA* mutations highlight an

indisputable connection with classic PD pathology, suggesting a common mechanism behind both familial and sporadic forms of PD.<sup>69</sup>

Hence, while *SNCA* mutations may cause rare and quite severe familial PD, it is clear that alpha-synuclein contained within LB is a common feature among all forms of PD, including both those with other familial mutations (i.e. *LRRK2*) and those of idiopathic etiology.<sup>62,69,99</sup> Given the correlation between *SNCA* copy number dosage with age of onset and severity of symptoms, a dose dependent relationship hypothesis has been suggested to occur between levels of alpha synuclein and severity of disease and there is some suggestion from GWA studies that synuclein levels are an important influence in typical PD.<sup>99</sup>

Alpha-synuclein protein forms dense fibrillar aggregates in LB inclusions, the pathognomonic hallmark of PD. Both *SNCA* CNVs and pathogenic point mutations enhance alpha-synuclein's transformation into an aggregated beta pleated sheet from its previous monomeric form. While transitioning into its new secondary protein structure, alpha-synuclein forms oligomer and fibrillar intermediates which are presumably pathogenic to neuronal cells in the SNPC.<sup>100,101</sup> Remarkably, in vivo investigations have revealed alpha-synuclein's ability to transmit its pathogenic secondary structure to neighboring cells in a prion-like propagation mechanism.<sup>102</sup> In addition, alpha-synuclein plays a critical role in maintaining the membrane curvature of the presynaptic terminal, which serves as an important site for neurotransmitter uptake and release. Thus, disruptions in normal alpha-synuclein function can have widespread effects on synaptic transmission.<sup>103,104</sup>

#### 1.4.2.1.3 *VPS35*

One of the more recent PD genes exhibiting autosomal dominant inheritance, vacuolar protein sorting 35 homolog (*VPS35*), was discovered in 2012 by exome sequencing.<sup>58,59</sup> With a frequency even lower than *SNCA* mutation carriers, those harboring disease-causing *VPS35* mutations were initially estimated to characterize 0.1% of the overall PD population.<sup>65</sup> Subsequent analyses, however, suggest that the *VPS35* p.D620N variant exists in approximately 1% of all familial PD cases with a widespread global distribution.<sup>105,106</sup> Analogous to the *LRRK2* p.G2019S mutation, *VPS35* p.D620N is seen among sporadic PD cases and similarly demonstrates variable penetrance.<sup>107</sup> Despite a slightly younger age of onset, the clinical presentation of *VPS35* mutation carriers resembles individuals with classic late onset, levodopa-responsive PD.<sup>65</sup>

Encoding a subunit of the retromer cargo recognition complex, *VPS35* serves as a key player in endosomal-lysosomal trafficking.<sup>65</sup> Specifically, communication between sorting nexins, the WASH complex and the retromer complex modulate the ability of transmembrane proteins to travel between endosomes, the trans Golgi network and the plasma membrane.<sup>108</sup> In cellular models overexpressing a pathogenic *VPS35* mutation, the retromer-WASH interaction destabilizes and hinders normal autophagosome formation and removal.<sup>109</sup>

#### 1.4.2.1.4 *ATXN2* and *ATXN3*

While mutations in *ATXN2* and *ATXN3* reside under the umbrella of spinocerebellar ataxias (*SCA*), characterized by decline in balance and coordination, patients may present with parkinsonian features.<sup>69</sup> Mutations in *ATXN2* are attributed to

CAG repeat expansions, with levels near the 34 repeat threshold most frequently observed in patients displaying parkinsonian phenotypes.<sup>69</sup> Likewise, in patients with triplet repeat expansions in *ATXN3*, responsible for Machado Joseph Disease (MJD), clinical presentation may consist of parkinsonism in addition to atypical features like neuropathy.<sup>69,110</sup>

#### 1.4.2.1.5 MAPT

Both splice site and missense mutations in *MAPT* have been attributed to causing Pick's Disease (FTDP-17), which consists of both Frontotemporal dementia with parkinsonism.<sup>111</sup> While parkinsonian features have been reported in the early stages of disease, neuropathology demonstrates the presence of tau, as opposed to alpha-synuclein, localized within neuronal and glial inclusions.<sup>111</sup> Though the absence of LB does not exclude PD in the differential diagnosis, it suggests pathogenic *MAPT* mutations are not responsible for typical PD and that it is likely that the pathological mechanism underlying FTDP-17 is distinct from that in typical PD.

#### 1.4.2.1.6 DCTN1

Resonating with deleterious mutations in *MAPT*, those in Dynactin subunit 1 (*DCTN1*) result in predominately tau inclusions infiltrating post-mortem brain tissue.<sup>112</sup> However, patients with mutations in *DCTN1*, responsible for Perry Syndrome, also exhibit rare neuropathological features like TAR DNA binding protein 43 (TDP-43) deposition.<sup>113</sup> While parkinsonism is a key malady of Perry Syndrome, it is usually preceded by severe neuropsychiatric symptoms including depression and significant weight loss, and later accompanied by respiratory failure.<sup>114</sup>

#### 1.4.2.1.7 GCH-1

In 2006 an individual with a pathogenic mutation in GTP cyclohydrolase 1 (*GTP-I*) was reported to exhibit both dystonia and parkinsonism.<sup>115</sup> It was hypothesized that this patient was presenting with a varied form of dopa-responsive dystonia (DRD) or perhaps even suffering from two distinct movement disorders, DRD and early-onset PD.<sup>115</sup> However, given the rarity of this mutation and corresponding phenotype, a definitive clinical diagnosis has yet to be determined.

### 1.4.2.2 Autosomal recessive PD

#### 1.4.2.2.1 PARK2

Initially identified by linkage analysis, pathogenic *PARK2/Parkin* mutations are diverse in nature, consisting of homozygous and compound heterozygous point mutations, as well as exonic deletions and duplications.<sup>62,65</sup> While most of these have been identified in familial cases, some have also been reported in idiopathic PD cases.<sup>62</sup> Interestingly, several individuals developing late onset PD have been shown to harbor heterozygous mutations in *PARK2*. However, as these variants have also been observed in healthy controls, their influence on disease development remains unknown.<sup>69</sup>

Patients carrying pathogenic *PARK2* mutations present with early onset PD, often before 45 years of age, with a minority of these (~4%) exhibiting signs before the age of 20, characterized as juvenile PD.<sup>64</sup> While *PARK2* mutations usually manifest a strong and consistent response to levodopa treatment, motor dysfunction progressively declines in patients at a young age.<sup>62</sup> Cases with a more advanced age of onset may display more atypical features including dystonia and hyperreflexia.<sup>116</sup>

The pathology of *PARK2* brain tissue is unique in the fact that LB are usually absent<sup>62</sup>; nonetheless, cases have been reported describing neurofibrillary tangles with the presence of LB in the substantia nigra and locus coeruleus, along with immunopositive alpha-synuclein inclusion bodies localized within the pedunculopontine nucleus.<sup>117–119</sup>

#### 1.4.2.2.2 *PINK1*

Representing only 8.4% of autosomal recessive familial PD cases and 3.7% of early onset PD cases including both sporadic and familial forms, *PINK1* mutations typically affect individuals in the 4<sup>th</sup> and 5<sup>th</sup> decades of life.<sup>120</sup> Discovered by homozygosity mapping among familial kindreds, *PINK1*, like *PARK2*, requires two mutated copies to cause disease and also exhibits a positive and sustained response to levodopa.<sup>121</sup> On the contrary, *PINK1* mutations may present with several atypical features including pyramidal signs, marked dystonia and sleep disturbances.<sup>122–124</sup>

Given the rarity of *PINK1* mutations, only a number of brains have been available for comprehensive post-mortem examination. While autopsy results have been largely heterogeneous, some have noted the classic PD features of SNPC neuronal loss and LB infiltration in the brainstem, SNPC and Meynert nucleus.<sup>69,125</sup> However, to fully elucidate *PINK1* neuropathological features will require the analysis of several more post-mortem brain specimens.

#### 1.4.2.2.3 *DJ-1/PARK7*

Identified by homozygosity mapping and positional cloning, *DJ-1* mutations characterize roughly 0.8% of familial and 0.4% of sporadic PD cases.<sup>120</sup> Typical age of onset is younger than those harboring *PARK2* or *PINK1* mutations, usually in the 2<sup>nd</sup> or



3<sup>rd</sup> decade of life.<sup>69</sup> While patients with *DJ-1* mutations are often highly responsive to levodopa, they may present with atypical features such as dysarthria and myoclonic jerks.<sup>126</sup>

#### 1.4.2.2.4 ATP13A2

Also known as Kufor-Rakeb (KR) syndrome, mutations in *ATPase* type 13A2 have been demonstrated to cause a juvenile onset parkinsonism using several genomic technologies: linkage analysis, homozygosity mapping and positional cloning.<sup>107,127</sup> Patients suffering from KR syndrome are affected as early as 12-15 years of age, often exhibiting rapid disease progression and decline with accompanying pyramidal symptoms.<sup>69,107</sup> Further, some cases have been noted to possess gross neurological deficits in conjunction with global CNS axonal loss.<sup>128,129</sup> However, given the few number of cases reported, clinical phenotype of KR syndrome has been shown to be variable in both disease progression and severity.<sup>69,127,128,130</sup> Finally, while neuropathological examinations have been limited, there appears to be a correlation between cases harboring LB inclusions with decreased *ATP13A2* levels, suggesting a common mechanism with typical forms of PD.<sup>128,131</sup>

*ATP13A2* is known to code for a transmembrane protein functioning in proteosomal degradation and lysosomal trafficking.<sup>62,132</sup> As lysosomal trafficking is necessary for normal mitochondrial regulation, including the lysosome-mediated removal of autophagosomes as well as acidification and stability of the lysosomal membrane, disruption of lysosomal function also inhibits healthy mitochondrial activity.<sup>133</sup> Transgenic models have illustrated that truncating mutations in *ATP13A2* led to cell preservation of defective *ATP13A2* protein and subsequent destruction in the

endoplasmic reticulum and proteasome.<sup>132</sup> Moreover, such models harboring homozygous *ATP13A2* loss of function mutations demonstrated the misfolding of alpha-synuclein and cell toxicity, providing further support for our evolving pathophysiological framework.<sup>62,132</sup>

#### 1.4.2.2.5 FBXO7

Similar to patients with *ATP13A2* mutations, those presenting with pathogenic mutations in F-box only protein 7, *FBXO7*, demonstrate a child-onset atypical parkinsonism accompanied by dystonia, pyramidal signs and equinovarus deformity.<sup>62,69</sup> Patients may present with psychological maladies, blepharospasm and symptoms of dyskinesia despite an initial positive response to levodopa.<sup>69</sup> Identified via linkage analysis, *FBXO7* cases are rare and post-mortem examination has been limited.<sup>134</sup>

While its primary role is largely unknown, *FBXO7* has demonstrated neuronal functions, such as synapse formation and cellular proliferation, through its association with the ubiquitin proteasome pathway.<sup>62</sup>

#### 1.4.2.2.6 PLA2G6

*PLA2G6* mutations discovered through homozygosity mapping have demonstrated extraordinary heterogeneity, consisting of both infantile onset forms as well as adult onset forms existing under the spectrum of neurodegeneration with iron accumulation (NBIA) disorders.<sup>135</sup> Characterized as NBIA type 2, *PLA2G6* is responsible for infantile neuroaxonal dystrophy (INAD), which presents before five years of age, coinciding with ataxia, dysarthria, dystonia, rigidity and developmental delays.<sup>69</sup> Categorically, classical INAD presents in patients below two years of age and progresses at a slow pace. In contrast, juvenile forms describe individuals affected between 2 and 18

years of age while those older than 18 are diagnosed with adult onset NAD or atypical NAD.<sup>69,128,135,136</sup>

While the clinical phenotype is variable, several cases have been reported with spasticity, seizures and optic nerve pallor.<sup>69</sup> Interestingly, both INAD and NAD forms demonstrate the same neuropathological signature: the presence of LB inclusions among all homozygous *PLA2G6* carriers, despite the fact that parkinsonism is not observed in all patients.<sup>131</sup> Furthermore, LB inclusions are surrounded by alpha-synuclein positive dystrophic neurites within both the substantia nigra and cortex, alongside neurofibrillary tangles displaying tau immune-reactivity.<sup>69,128</sup>

Encoding the catalytic enzyme calcium independent phospholipase A2, *PLA2G6* functions in forming free fatty acids, which regulate apoptosis and inflammation.<sup>137,138</sup>

#### 1.4.2.2.7 PANK2

Known as NBIA type I, mutations in pantothenate kinase 2 (*PANK2*) typically present within the first or second decade of life, notably with rapid disease progression, including the inability to ambulate a couple of years later and fatality shortly thereafter.<sup>139</sup> Patients have been reported to demonstrate a variety of gait disturbances, including clumsiness and imbalance, as well as hand tremor dysarthria and cognitive dysfunction.<sup>69</sup> While such features denote typical symptoms of patients with *PANK2* mutations, others have exhibited supranuclear vertical gaze palsy and facial hypomania.<sup>69</sup> Moreover, atypical phenotypic forms of *PANK2* mutation may also display extrapyramidal features but often maintain ambulatory function. In both typical and atypical *PANK2* mutations, levodopa responsiveness is usually positive but declines within a 1-2 year period.<sup>139</sup>

Brains harboring *PANK2* mutations reveal a classic “eye-of-the-tiger” sign, consisting of hypointensive regions of iron deposition in peripheral locations of the globus pallidus, surrounding regions of hyperintensity, presumably due to gliosis, in the central globus pallidus.<sup>140</sup> Post-mortem brain tissue characterized by *PANK2* mutations lack both LB inclusions and alpha-synuclein positive dystrophic neurites, a marked distinct from autopsy results of *PLA2G6* mutation carriers.<sup>141</sup> While neurofibrillary tangles of tau immunoreactivity may be present, there is no pathognomonic hallmark for *PANK2* neuropathology.<sup>141</sup>

#### 1.4.2.2.8 Other rare recessive forms of PD

Several other case reports have suggested new genes that may be involved in recessive forms of atypical PD. This includes both *DNAJC6* and *SYNJ1*, both of which were recently identified through WES and homozygosity mapping.<sup>142–144</sup> Interestingly, an X-linked recessive gene, *ATP6AP2*, has been declared a candidate of atypical PD upon discovery via WES and linkage analysis.<sup>145</sup> While part of the spastic paraplegia family, *SPG11* harbors mutations that have demonstrated features of atypical PD.<sup>128,135,146,147</sup> A unifying theme among *SPG11*, *PLA2G6* and *FBXO7* is that they have all demonstrated brain iron accumulation and supranuclear gaze palsy, despite only *PLA2G6* characterized as a disease of iron accumulation.<sup>107</sup> Lastly, parkinsonism has been reported in the clinical phenotype of patients with mutations in fatty acid 2-hydroxylase (*FA2H*) and alpha chain of type XVIII collagen (*COL18A2*), though the pathophysiological etiology resulting in PD is unknown.<sup>135,148,149</sup>

### 1.4.3 Molecular mechanisms of PD gene mutations

Given the large number of PD genes that have been identified in the last two decades, significant effort has been placed in the characterization of gene function through cell and transgenic work. As detailed above a few common molecular processes have been suggested as critical in PD pathophysiology: Endosomal protein sorting and recycling, lysosome mediated autophagy, synaptic transmission, and mitochondrial quality control.<sup>65</sup> To unravel the molecular mechanisms driving these cellular functions, the study of individual genes and their interactions in a pathway-based analysis has been insightful.

A goal in much of this work has been to unite the proteins encoded by PD-linked genes into a common pathway. Perhaps the most success has been had in this regard within the autosomal recessive genes. The proteins encoded by *PINK1*, *PARK2* and *DJ-1*, share a common cellular mechanism: mitochondrial quality control and regulation.<sup>131</sup> Broadly, this includes mitogenesis, mitophagy, mitochondrial homeostasis and transport.<sup>62</sup>

In healthy mitochondria, PINK1 protein resides on the inner mitochondrial membrane, normally undergoing cleavage and traveling to the cytoplasm.<sup>150</sup> However, this process is disrupted upon reduction of the membrane potential, causing *PINK1* to bind to the outer mitochondrial membrane.

During depolarization, *Parkin*, an E3 ubiquitin ligase, is recruited to the membrane and consequently phosphorylated by *PINK1*. The latter step results in several effects: inhibition of mitofusion via the ubiquitination of mitofusion, dysregulation of mitochondrial trafficking via ubiquitination of the Miro/Milton complex, and finally the

loss of a key mitophagy signal through Voltage dependent anion channel 1 (VDAC1) ubiquitination.<sup>151–158</sup> These series of events result in extensive accumulation of damaged, bio-energetically comprised mitochondria and ultimately mitochondrial dysfunction.<sup>154,157,159</sup>

In addition to these functions, *Parkin* has demonstrated a role in maintaining mitochondrial biogenesis through an alternative pathway via interaction with PARIS/PGC1 $\alpha$ .<sup>160</sup> Likewise, *PINK1* also portrays an additional role in mitochondrial homeostasis through regulation of calcium levels.<sup>161</sup> Lastly, while the role of Daisuke-Junko-1 (*DJ-1*) continues to be unraveled, it functions as a powerful antioxidant which migrates across the mitochondrial membrane, possessing a presumably neuroprotective function.<sup>162</sup> Thus, pathogenic mutations in these three genes all result in dysfunctional mitochondria.

#### **1.4.4 Integrating critical molecular processes regulated by PD proteins**

While researchers continue to elucidate the specific molecular processes underlying PD causing disease genes, Trinh et. al illustrates the integration of these puzzle pieces. Among the first of these core processes includes synaptic transmission, encompassing both endocytosis and exocytosis, in conjunction with endosomal receptor sorting and recycling.<sup>64</sup>

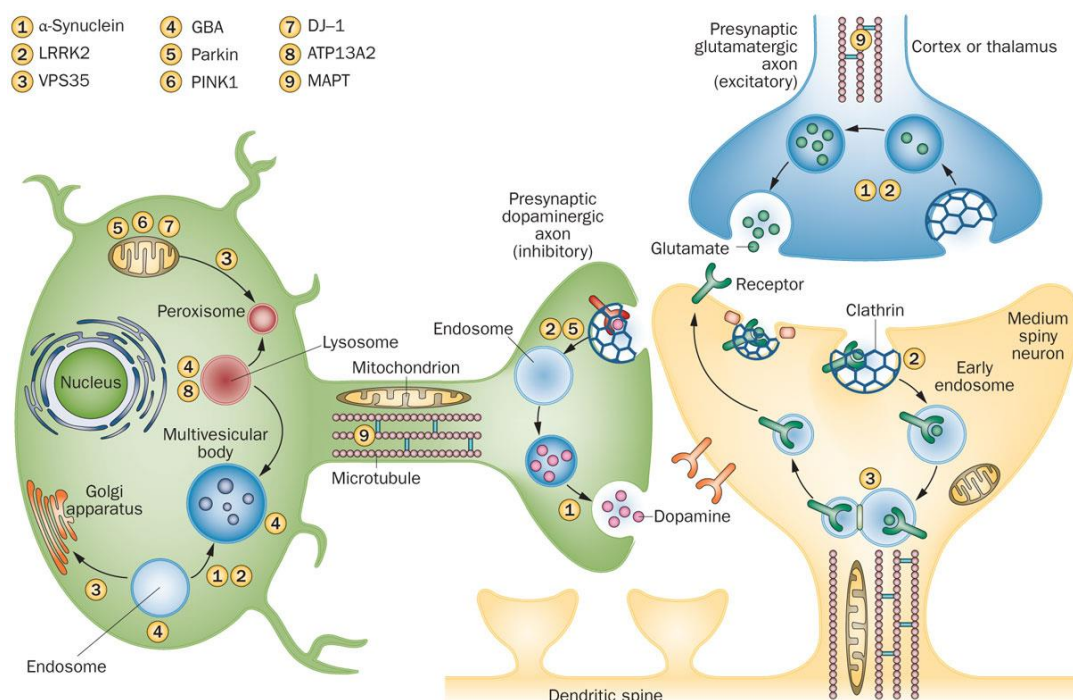
As portrayed in Figure 10, *SNCA* (1) facilitates exocytosis and also maintains a role in the process of endocytosis in the presynaptic glutamatergic nerve terminals within the cortex and thalamus. Located at the postsynaptic dopaminergic terminals, *LRRK2* (2) controls the phosphorylation of endophilin A while regulating the release of

clathrin-coated endocytotic vesicles. Further, acting in the presynaptic glutamatergic and medium spiny neurons, *LRRK2* is responsible for phosphorylation of *MAPT* (9), stabilization of microtubules, polarity and branching of neurons, as well as chaperone-assisted autophagy. Upon phosphorylation, *MAPT* mediates axonal cargo trafficking and delivery. The third known gene responsible for autosomal dominant PD, *VPS35* (3), assists in regulation of early endosome cargo identification and the membrane recruitment process; subsequently, this creates a clathrin-independent carrier in medium spiny neurons. Moreover, *VPS35* regulates the recycling that occurs between endosomes and either the Golgi apparatus or plasma membrane in the cell body of dopaminergic neurons of the substantia nigra. Lastly, vesicle movement between the mitochondria and peroxisomes is under surveillance by *VPS35*.<sup>64</sup>

*GBA* (4), which will be discussed in greater detail in the upcoming risk genes section, also resides in the dopaminergic neuron of the substantia nigra and utilizes the retromer for receptor recycling. As a lysosomal acid hydrolase, *GBA* plays a fundamental role in the next key process in PD pathogenesis: lysosome-mediated autophagy.

Likewise, *ATP13A2* (8) is closely associated as well.

The unifying mitochondrial quality control pathway, mediated via *Parkin* (5), *PINK1* (6), and *DJ-1* (7), is illustrated in the dopaminergic neuron of the substantia nigra. Specifically, while *Parkin* aids in ubiquitination and proteasomal processes, *PINK1* joins *Parkin* to assist in mitochondrial stability. These two proteins, in concert with *DJ-1*, are essential for healthy mitochondrial biogenesis and the initiation of autophagy.<sup>64</sup>



**Figure 10: Putative molecular mechanisms underlying PD**

Genes harboring variants responsible for monogenic forms of PD or associated with variants that elevate PD risk are labeled numerically. Each gene interacts with different aspects of the presynaptic dopaminergic axon, presynaptic glutamatergic axon, and/or medium spiny neurons to carry-out key cellular functions utilizing neurotransmitters, organelles, and transport associated molecules (i.e. endosomes).

(Reproduced from Trinh et. al. 2013).<sup>64</sup>

Functional analysis of known PD genes associated with early onset familial forms and late onset (familial or sporadic) forms have revealed an interesting dichotomy of molecular mechanisms: while dysfunction of synaptic transmission and vesicular recycling are strongly related to late-onset PD pathophysiology, early forms are associated with mitochondrial dysfunction and lysosomal degradation.<sup>65</sup>

Using a pathways-based approach, which investigates protein networks among PD associated genes, protein-protein interactions can be identified and incorporated into our framework of PD etiology. Notably, cells expressing knockout *GBA* mutations



associated with PD have demonstrated enhanced *SNCA* accumulation.<sup>163,164</sup> Specifically, the accumulation of glucocerebrosidase substrate glucosyl ceramide has been suggested to stabilize alpha-synuclein fibril formation within lysosomes.<sup>165</sup> Similarly, *SNCA* overexpression causes a decrease in *GBA* activity, suggesting that a *SNCA* “priming effect” may be a prerequisite for the pathogenicity of *GBA*.<sup>131,165</sup> Once this vicious cycle is established, the notion of a possible “bidirectional feedback loop,” in which accumulated alpha-synuclein fibrils continuously inhibit normal *GBA* trafficking to the lysosome, may persist indefinitely.<sup>107,165</sup> While hypothetical, this feedback mechanism has been suggested to account for the accelerated speed at which cortical synuclein pathology develops in heterozygous *GBA* mutation carriers.<sup>107</sup>

Further investigation of mitochondrial dysfunction has been pursued via analysis of mitochondrial function among post mortem brain tissue in PD patients. One study found that the substantia nigra of PD brains demonstrated mitochondrial complex I inhibition.<sup>131,166</sup> Maternally inherited homoplasmic mutations in mitochondrial DNA, which are regulated exogenously by nuclear DNA and endogenously by mitochondrial DNA (mtDNA), have been found to be elevated among pathologically examined PD brains. Further, there may be an association between risk for sporadic PD and mitochondrial haplotype.<sup>167,168</sup>

#### **1.4.5 Risk loci in PD**

The number of genes and loci demonstrating statistical significance on GWA studies continues to expand, particularly upon subsequent meta-analyses deriving greater statistical power.<sup>169</sup> We have learned that several of the PD disease genes that have been

deemed causal actually harbor many common variants (MAF >5%) that enhance one's risk toward developing PD.

Prior to the advent of genome wide association methodologies the principal tool to test association was the candidate gene association test. Typically this involved testing variability in a gene of functional relevance for association with disease. This included, most obviously, genes that had been previously shown to contain disease-causing mutations. While by and large the low resolution and poor power of these studies meant that the results were generally unreliable, there are two notable exceptions in the context of PD.

*SNCA* mutations were linked to PD in 1997 and over subsequent years a large number of manuscripts were published that tested for association between common variants at this gene and risk for typical PD. The majority of these studies centered on the *SNCA* REP1 variant, which is approximately 10kb 5' to the transcription start site of *SNCA*. The results of these experiments were quite mixed, however, in 2006 a meta-analysis established clear association between this variant and risk for disease. Notably this work required 2692 cases and 2652 controls.<sup>170</sup> *In vivo* cell work has suggested that genetic variation in this region influences gene transcription and possibly gene expression, although this functional consequence has been called into question more recently<sup>64,170–172</sup>.

As with *SNCA*, *LRRK2* was examined for association with common apparently sporadic disease, outside of rare monogenic mutations. Notably, initial candidate association based work identified a variant particular to Asian populations, p.G2385R,

which roughly doubles one's risk of developing PD, being carried by ~3% of the general population and ~6% of the PD population <sup>173,174</sup>.

While these two early examples of genetic association in PD resulted from originally unbiased linkages of these genes to monogenic forms of PD, the third example of successful association was a result of astute clinical observation. Initially, clinicians noticed that several patients with Gaucher's Disease demonstrated a parkinsonian phenotype. Further clinical observation revealed that first and second-degree relatives of individuals with Gaucher's disease manifested an increased incidence of PD.<sup>175,176</sup>

Since bi-allelic variants of *GBA* causes Gaucher's disease, a lysosomal storage disorder characterized by glucosylceramide accumulation, variants in *GBA* were scrutinized for an association with PD.<sup>176</sup> Notably, a meta analysis conducted by Sidransky et al. revealed that heterozygous *GBA* mutations of this very same variant are the largest genetic risk factor for developing PD, enhancing one's risk approximately five-fold.<sup>7,176</sup>

While there is vast population heterogeneity, it is estimated that this variant occurs in approximately 1% of the global population.<sup>62</sup> Moreover, variation in this *GBA* allele also adheres to concept of variable penetrance in an age-dependent fashion and has been shown to be a significant risk factor for DLB with and without AD pathology, suggesting a possible shared mechanism underlying cerebral LB inclusions.<sup>62,177</sup>

#### ***1.4.5.1 Risk loci identified by GWA***

In 2005, the very first GWA study on PD was performed using 195 cases from the United Kingdom and genotyping 5546 microsatellite markers in these samples. Replication of original findings within an independent cohort failed to reveal any

significant associations for cases of apparently sporadic PD, while a single marker, DIS2886, manifested an association with familial PD.<sup>178</sup> Just a year later, Fung et al performed genome-wide genotyping on 267 cases and 270 controls, using more than 408,000 genotypic markers and made this data publically accessible to the scientific community.<sup>179</sup>

Several subsequent PD GWA studies have been performed within the last decade, with a trend of increasing sample size, microsatellite markers, and overall statistical power. Many of these studies continued to reveal new loci while also confirming those previously identified.<sup>37,180</sup> Furthermore, larger and more homogenous sample cohorts have facilitated the identification of PD associated variants within distinct ethnic cohorts.<sup>181,182</sup>

In 2014, Nalls et al performed the largest GWA meta analysis to date, genotyping almost 8 million SNPs in an impressive 13,808 PD cases and 95,282 controls. A total of 28 loci were deemed statistically significant, 6 of which had been identified in prior PD GWA studies.<sup>169</sup> Many of these loci have demonstrated an association with both familial and apparently sporadic forms of PD (Table 2).

SNP Information							Discovery phase (13,728 cases and 95,282 controls)		Replication phase (5,353 cases and 5,551 controls)		Joint phase (19,081 cases and 100,833 controls)	
SNP	C	Position (bp)	Nearest gene(s)	Effect allele	Alternate allele	Effect allele frequency	Odds ratio	P	Odds ratio	P	Odds ratio	P
Genome Wide Significant, Discovery Phase												
rs35749011*	1	155,135,036	GBA/SYT11	a	g	0.017	1.762	6.09×10 <sup>-23</sup>	2.307	7.48×10 <sup>-09</sup>	1.824	1.37×-29
rs823118	1	205,723,572	RAB7L1/NUCKS1	t	c	0.559	1.126	1.36×10 <sup>-13</sup>	1.109	1.43×10 <sup>-04</sup>	1.122	1.66×-16
rs10797576	1	232,664,611	SIPA1L2	t	c	0.14	1.139	1.19×10 <sup>-08</sup>	1.11	3.38×10 <sup>-03</sup>	1.131	4.87×-10
rs6430538	2	135,539,967	ACMSD/TMEM163	t	c	0.43	0.873	5.56×10 <sup>-15</sup>	0.882	9.42×10 <sup>-06</sup>	0.875	9.13×-20
rs1474055*	2	169,110,394	STK39	t	c	0.128	1.213	7.12×10 <sup>-16</sup>	1.218	1.07×10 <sup>-06</sup>	1.214	1.15×-20
rs115185635*	3	87,520,857	KRT8P2/APOOP2	c	g	0.035	1.789	2.18×10 <sup>-08</sup>	0.931	0.846	1.142	0.022
rs12637471	3	182,762,437	MCCC1	a	g	0.193	0.844	3.32×10 <sup>-16</sup>	0.836	3.72×10 <sup>-07</sup>	0.842	2.14×-21
rs34311866	4	951,947	TMEM175/GAK/DGKQ	t	c	0.809	0.784	3.58×10 <sup>-33</sup>	0.791	6.29×10 <sup>-12</sup>	0.786	1.02×-43
rs11724635	4	15,737,101	BST1	a	c	0.553	1.122	8.07×10 <sup>-13</sup>	1.138	2.73×10 <sup>-06</sup>	1.126	9.44×-18
rs6812193	4	77,198,986	FAM47E/SCARB2	t	c	0.364	0.897	7.17×10 <sup>-11</sup>	0.935	0.011	0.907	2.95×-11
rs356182	4	90,626,111	SNCA	a	g	0.633	0.737	3.23×10 <sup>-67</sup>	0.822	1.75×10 <sup>-12</sup>	0.760	4.16×-13
rs9275326*	6	32,666,660	HLA-DQB1	t	c	0.094	0.797	5.82×10 <sup>-13</sup>	0.9	0.018	0.826	1.19×-12
rs199347	7	23,293,746	GPNMB	a	g	0.59	1.123	2.37×10 <sup>-12</sup>	1.072	7.66×10 <sup>-03</sup>	1.110	1.18×-12
rs117896735*	10	121,536,327	INPP5F	a	g	0.014	1.767	1.21×10 <sup>-11</sup>	1.404	1.10×10 <sup>-03</sup>	1.624	4.34×-13
rs3793947*	11	83,544,472	DLG2	a	g	0.443	0.912	2.59×10 <sup>-08</sup>	0.976	0.201	0.929	3.96×-07
rs329648	11	133,765,367	MIR4697	t	c	0.354	1.1	1.65×10 <sup>-08</sup>	1.121	4.38×10 <sup>-05</sup>	1.105	9.83×-12
rs76904798	12	40,614,434	LRRK2	t	c	0.143	1.17	1.33×10 <sup>-12</sup>	1.11	3.69×10 <sup>-03</sup>	1.155	5.24×-14
rs11060180	12	123,303,586	CCDC62	a	g	0.558	1.101	2.14×10 <sup>-08</sup>	1.114	7.26×10 <sup>-05</sup>	1.105	6.02×-12
rs11158026	14	55,348,869	GCH1	t	c	0.335	0.889	7.13×10 <sup>-11</sup>	0.948	0.039	0.904	5.85×-11
rs1555399*	14	67,984,370	TMEM229B	a	t	0.468	0.872	5.53×10 <sup>-16</sup>	0.971	0.144	0.897	6.63×-14
rs2414739	15	61,994,134	VPS13C	a	g	0.734	1.114	4.13×10 <sup>-09</sup>	1.109	7.96×10 <sup>-04</sup>	1.113	1.23×-11
rs14235	16	31,121,793	BCKDK/STX1B	a	g	0.381	1.094	3.89×10 <sup>-08</sup>	1.133	7.72×10 <sup>-06</sup>	1.103	2.43×-12
rs17649553	17	43,994,648	MAPT	t	c	0.226	0.771	4.86×10 <sup>-37</sup>	0.764	7.03×10 <sup>-15</sup>	0.769	2.37×-48
rs12456492	18	40,673,380	RIT2	a	g	0.693	0.905	5.12×10 <sup>-09</sup>	0.9	2.16×10 <sup>-04</sup>	0.904	7.74×-12
rs62120679*	19	2,363,319	SPPL2B	t	c	0.314	1.141	2.53×10 <sup>-09</sup>	0.999	0.518	1.097	5.57×-07
rs8118008*	20	3,168,166	DDRKG1	a	g	0.657	1.111	2.32×10 <sup>-08</sup>	1.113	1.18×10 <sup>-04</sup>	1.111	3.04×-11
Previously Reported as Significant in Genome Wide Studies												
rs34016896	3	160,992,864	NMD3	t	c	0.319	1.08	7.68×10 <sup>-06</sup>	1.028	0.174	1.067	1.08×-05
rs591323	8	16,697,091	FGF20	a	g	0.275	0.921	1.30×10 <sup>-05</sup>	0.902	6.16×10 <sup>-04</sup>	0.916	6.68×-08
rs60298754	8	89,373,041	MMP16	t	c	0.024	1.078	0.181	-	-	1.078	0.181
rs7077361	10	15,561,543	ITGA8	t	c	0.874	1.11	3.24×10 <sup>-05</sup>	1.044	0.154	1.092	4.16×-05
rs11868035	17	17,715,101	SREBF/RAI1	a	g	0.298	0.937	2.17×10 <sup>-04</sup>	0.947	0.036	0.939	5.98×-05
rs2823357	21	16,914,905	USP25	a	g	0.37	1.036	0.032	1.018	0.267	1.031	0.027

C - Chromosome; OR - odds ratio

\* replication genotyping for these SNPs failed assay design or quality control and a suitable proxy variant was selected (rs35749011, proxy rs71628662; rs1474055, proxy rs1955337; rs115185635, proxy rs62267708; rs117896735, proxy rs118117788; rs3793947, proxy rs12283611; rs1555399, proxy rs1077989; rs62120679, proxy rs10402629; rs8118008, proxy rs55785911). Note, only replication phase p-values are one-sided. Nearest gene or previously published proximal gene names included in table.

**Table 2: Results of PD GWA study**

Discovery and replication stages.  
(Reproduced from Nalls et al. 2014).<sup>169</sup>

Despite the fact that PD is member of the alpha-synucleinopathy family, the *MAPT* H1 haplotype, spanning 1.5M, is significantly more common in PD cases than controls among Caucasian populations.<sup>183,184</sup> Further, the snp rs242557 in the H1c region manifests a strong association with PSP, CBD and Parkinson-Dementia complex in

Guam.<sup>185–187</sup> This is noteworthy as the former two are tauopathies and previously thought to have a distinctive pathological signature from PD.<sup>188</sup>

Other genes associated with several lysosomal storage disorders, including *HEXA*, *MCOLN1* and *SMPD1*, causing Tay-Sachs, mucopolipidosis type IV and Niemann-Pick disease, respectively, have been tested for an association with PD based on *GBA* findings. While the former two genes failed to reveal a significant association with PD risk, a variant in *SMPD1* (p.L302P) was demonstrated to increase the risk of PD by a factor of nine in an Ashkenazi Jewish PD patient cohort.<sup>62,189</sup>

In addition to *LRRK2*, *SNCA*, *MAPT*, *GBA* and *SMPD1*, several other genes have exceeded statistical significant on large-scale GWA studies. The largest PD GWA meta-analysis to date, which combined SNP data from 15 different European GWA studies, revealed an impressive 28 variants among 24 loci manifesting an association with PD.<sup>169</sup> While many of these risk loci do not exhibit large effect sizes, risk variant pooling demonstrated a three-fold increase in PD risk among carriers residing in the highest risk quintile.<sup>65,169</sup>

Among the GWA hits, some of these genes play key roles in the immune system. Bone marrow stromal cell antigen 1 (*BST1*), for instance, is involved in neutrophil adhesion and migration.<sup>64</sup> The *HLA-DRA* and *HLA-DRB* loci, which code for MHC class II cell surface molecules, are involved in inflammation and autoimmune disease.<sup>107,190,191</sup>

Others have been suggested to function in some of the key integrated molecular processes underlying PD. For example, *RAB7L1*, located in the *PARK16* locus, interacts with both *LRRK2* and *VPS35*, with a possible role in endosomal-lysosomal trafficking.<sup>192</sup> *GAK*, located within the *GAK-DGKQA* locus, is expressed by *DNAJC6*, playing a role in

clathrin-mediated endocytosis. Familial PD GWA results identified *DGKQ* and phosphatidylinositol kinase (*PIK3CD*) as significant hits, both of which are critical in regulating membrane curvature and signal transmission.<sup>193,194</sup>

Finally, the function of other significant GWA hits is a subject of ongoing investigation. Mutations in *GCHI*, as discussed previously, can cause DRD in childhood. Interestingly, individuals carrying mutations in *GCHI* have a seven-fold increased risk of developing adult forms of idiopathic PD.<sup>65,195</sup>

#### 1.4.6 Interpretation of GWA findings and PD etiology

GWA studies have demonstrated to be a very powerful tool for not only identifying new genes and loci associated with PD risk but additionally confirming known disease-causing genes. Studying monozygotic and dizygotic twins revealed concordance values ranging from 11-15.5% and 4-11%, respectively, which are vastly distinct from fully penetrant Mendelian diseases.<sup>69,196</sup> Further, a PD heritability analysis based on common genetic variants in GWA studies was also quite insightful, as we learned that approximately 27% of PD is heritable through common genetic variation.<sup>49</sup> Given that the heritability estimate is based on common variation alone, it does not account for many of the rare variants known to cause familial forms of PD. For instance, several of the *GBA* locus mutations, 17 of which are considered rare, are not included in the genotyping array.<sup>49</sup> Notably, the total variance accounted for by GWA study SNPs in PD is estimated to be 6-7%, highlighting the importance of looking for rare PD associated variants beyond the scope of those genotyped.<sup>49,197</sup>

While furthering our understanding of genes associated with both familial and sporadic forms of PD is integral for establishing functional pathways and molecular

mechanisms of disease, this suggests that PD genetic etiology lies on a continuum, ranging from a classical Mendelian inheritance of rare variants to graded levels of risk from common variants in those same genes. Such multifactorial inheritance patterns among several genes (*SNCA*, *LRRK2*) suggest the plausible nature of epistasis, whereby such genes may interact to contribute in both sporadic and familial forms of PD.<sup>2,7,198</sup>

From the results of these studies, we can draw two important conclusions: First, while we know that PD is a complex polygenic disorder, there are still more genetic variants that have yet to be discovered. Second, while genetics clearly plays a pivotal role in PD risk and development, it does not explain the comprehensive PD landscape. Hence, we must also consider the full scope of scientifically quantifiable causes, including both epigenetic and environmental factors.

#### **1.4.7 Making progress in PD**

As we consider our current ability to diagnose, intervene and manage PD, from the prodromal phase to the late stages of disease, substantial work must be done. First, as we continue to acquire more information on genetic heterogeneity and mutation frequency of PD associated genes, we must continue to inform the international PD community to further our understanding of PD pathophysiology. This is feasible through a large-scale PD database, which would acquire evidence for putative causal genes through identification of individuals carrying mutations in heterogeneous cohorts.<sup>65</sup> An important caveat of this is that it remains strictly within the scientific community, for the clinical utility of genetic risk profiling information is hazy at best. Hence, when the FDA suspended the 23and me screening service of common disease variants, among them PD genes, this was in an effort to prevent the infiltration of genetic information lacking



corresponding genetic counseling or viable therapeutic options into the public sector (23andme).<sup>131,199</sup>

As it has been estimated that PD is roughly 27% heritable based on common variants in GWA studies, there are many risk genes that have yet to be identified.<sup>49</sup> While some of these are likely low risk alleles, requiring substantially large sample sizes for adequate power and detection, others may be harboring intermediate risk levels like *GBA*. Finally, the excess homozygosity among early onset PD cases lacking disease causal mutations in known PD associated genes suggests there are more autosomal recessive PD genes.<sup>107,200</sup>

## 1.5 Multiple system atrophy

MSA is a rare progressive neurodegenerative disease with an estimated incidence of 3-4 per every 100,000 individuals among adults 50-99 years of age, and is clinically defined by a triad of cerebellar ataxia, parkinsonism and autonomic dysfunction in conjunction with pyramidal signs.<sup>201–203</sup> From an average age of onset of 57, to mortality, MSA typically progresses over 7-9 years and affects both sexes equally.<sup>202,204</sup> However, with our limited understanding of the genetics and biomarkers of MSA, definite diagnosis can only be verified pathologically.<sup>202</sup> In addition to an estimated false positive clinical diagnostic rate of approximately 14%, MSA's clinical presentation is often not realized until later stages of disease progression, with very limited clinical ability to intervene.<sup>205,206</sup> For definitive diagnosis, one must validate the presence of MSA's unique histological hallmark: alpha-synuclein-positive glial cytoplasmic inclusions (GCIs).<sup>207</sup>

Among Caucasian populations, it has been suggested that specific polymorphisms of the *SNCA* gene have been associated with an elevated risk of MSA.<sup>208</sup> As an etiologically and clinically complex disorder, MSA has been split into subtypes based on predominant clinical features.<sup>207</sup> A handful of studies have described significant population-specific variation among MSA patients regarding predominance of one subtype over another. This notion further supports a role of genetic etiology associated with specific risk factors in the development of MSA pathogenesis.<sup>209,210</sup> While *in vitro*, *in vivo* and transgenic studies continue to elucidate molecular mechanisms driving MSA etiology and pathology, the genetic underpinnings of this disease still requires extensive investigation. We describe here the state of the field in MSA, and urge that it is essential to apply state-of-the-art genetic approaches to MSA.<sup>201</sup> Ultimately, understanding the molecular pathogenesis of this disease is our best opportunity to design and test etiologic based interventions.

### **1.5.1 Clinical and neuropathological features of MSA**

Despite that an autopsy is necessary for a diagnosis of MSA, clinical diagnosis is often sought at the time of initial presentation.<sup>202</sup> Essentially, this is based upon a thorough clinical evaluation, revealing motor dysfunction (either parkinsonism or cerebellar), and/or autonomic dysfunction (excluding erectile dysfunction). It is hypothesized that subclinical neuropathological alterations may occur years before patients become clinically symptomatic.<sup>202</sup>

As a member of the alpha-synucleinopathy family, defined by well-demarcated alpha-synuclein-immunoreactive inclusions and aggregation, MSA's clinical presentation delineates several overlapping features with other members including PD and dementia

with Lewy bodies (DLB).<sup>201</sup> In a very small study consisting of 33 MSA patients and 80 controls, an increased frequency of neurological symptoms among first-degree relatives was reported.<sup>211,212</sup> Further, 5 individuals among a cohort of 38 pathologically confirmed MSA samples were shown to have at least one first or second-degree relative with parkinsonism.<sup>212,213</sup> Nonetheless, given these extremely low cohort numbers lacking requisite statistical power, a positive family history of PD has not been demonstrated to be a significant risk factor for the development of MSA, an observation that is perhaps confounded by the difficulty of clinically diagnosing MSA.

While MSA predominately consists of GCIs containing alpha-synuclein aggregates, it is important to note that other protein aggregates, including hyperphosphorylated tau, can also be found.<sup>53</sup> Interestingly, MSA also delineates extensive clinical overlap with members of the tauopathy family, including progressive supranuclear palsy (PSP) and corticobasal degeneration (CBD).<sup>214</sup> In a similar fashion to MSA with fellow alpha-synucleinopathies like PD, pathologically confirmed cases of MSA, PSP and CBD, all of which are considered “atypical parkinsonisms,” often present with phenotypes distinct from their “classical ones”; hence, MSA can present with a spectrum of clinical phenotypes (i.e. vertical gaze palsy), usually associated with tauopathies.<sup>214,215</sup> To address this uncertainty, studies have scrutinized cases of atypical parkinsonisms to establish well-defined criteria to increase diagnostic accuracy in a clinical context.<sup>214,215</sup>

In addition to clinical features of alpha-synucleinopathies and tauopathies, MSA phenotypes can also resonate with subtypes of spinocerebellar ataxias (SCAs) and other familial ataxias.<sup>216,217</sup> While the majority of SCAs are alpha-synuclein negative upon

immunohistochemical staining, a few subtypes, such as *SCA3*, can exhibit glial alpha-synuclein-positive inclusions.<sup>216,217</sup> Notably, some cases of *SCA3* can manifest features that are highly characteristic of MSA, including but not limited to: levodopa-responsive parkinsonism, pyramidal tract dysfunction and even some dysautonomia. Furthermore, substantial clinical overlap may exist between MSA and other genetic forms of SCA, including *SCA2*, *SCA6*, *SCA8*, and *SCA17*. In a cohort of 302 clinically diagnosed MSA patients, 7.3% were found to be SCA positive, of which more than half were *SCA17* carriers<sup>217–224</sup>. When MSA is in the differential diagnosis, it is recommended to perform genetic testing for the spinocerebellar ataxias in such patients to essentially rule out a familial ataxia.<sup>224,225</sup>

Based on pathological studies of regions predominately affected and their corresponding phenotypes, MSA has been subdivided into two distinct subtypes: MSA-Cerebellar (C), MSA-parkinsonism (P), with the prevalence varying in a population-specific manner.<sup>201,225</sup> Despite this clearly defined classification system of MSA, current treatment options for patients with either subtype are far from ideal: while there is no therapy to delay the progression of disease, levodopa is considered the primary treatment for symptoms, which exhibits a “modest and non-sustained effect.”<sup>226,227</sup> Despite that approximately 30% of MSA patients manifest an initial response to levodopa therapy, the response does not persist yet patients often find it challenging to wean themselves off of this drug.<sup>228</sup>

While MSA is considered an oligodendroglipathy with the pathological hallmark of widespread alpha-synuclein-immunoreactive GCIs (Papp-Lantos inclusions), MSA patients exhibit marked neurodegenerative changes in the striatonigral and/or

olivopontocerebellar structures of the brain.<sup>207</sup> There is a vast degree of variation in the degeneration, depicted by a broad spectrum of myelin pallor, gliosis and neuronal loss; nonetheless, such features are classic neuropathological manifestations of all MSA subtypes.<sup>201</sup> When differentiating MSA from CBD and PSP, gross differences in size and pallor of affected regions can provide valuable information.<sup>201</sup> Moreover, several case reports have documented the coexistence of tau and alpha-synuclein inclusion bodies within autopsies of a single individual, suggesting a common pathological mechanism, potentially through disruption of cytoskeletons and dislocation and aggregation of various proteins.<sup>202,229</sup>

It has been suggested that specific MSA clinical subtypes, duration of disease, and disease severity are all associated with the quantitative distribution and density of GCIs in MSA cases.<sup>201</sup> While GCIs represent the pathological signature of MSA, the abnormal accumulation of alpha-synuclein has also been identified within neuronal cytoplasmic inclusions (NCIs), neuronal nuclei inclusions (NNIs), and within neurites of a minority of MSA affected brains. While these findings have not been the primary focus of MSA in previous molecular research, the potential role of NCIs, NNIs and neurites in the pathological process of MSA has warranted further investigation.<sup>201,230</sup>

Within the last year, Cykowski et al embarked on an extensive neuropathological investigation of MSA post-mortem brains and revealed that widespread neuronal inclusions were seen in most patients, in both disease-associated regions (i.e. substantia nigra), and several other non-disease associated regions (i.e. hypothalamus). Further, a hierarchal region specific susceptibility pattern was observed from neuronal inclusions. While this was unrelated to clinical phenotype, the severity of pathology was disease

duration dependent. Moreover, interregional correlations between pathological neuronal and glial lesion burden were observed, hinting at possible overlapping disease mechanisms in distinct brain regions and the significance of NCIs and NNIs in MSA histopathology.<sup>231</sup>

A recent study investigated the immunohistochemistry underlying Minimal change MSA (MC-MSA), in which MC-MSA is defined as a subtype of MSA manifesting neuronal loss primarily in the substantia nigra and locus coeruleus.<sup>232</sup> Ling et al identified a greater proportion of NCIs in the disease-associated regions (substantia nigra, caudate) of MC-MSA individuals than in MSA controls. As neuronal changes were demonstrated to be disease duration dependent by Cykowski and colleagues, this suggests that NCIs may be involved early in the disease process. Collectively these findings suggest that alpha-synuclein associated oligodendroglial pathology (i.e. GCIs) could result or possibly occur in parallel with neuronal dysfunction (i.e. NCIs) capable of causing clinical symptoms prior to neuron loss.<sup>231</sup>

Corresponding with the clinically defined subtypes of MSA, gross pathological depictions of cerebellar and parkinsonian subtypes parallel those same regions or systems predominantly affected by MSA pathology.<sup>201</sup> In cases of MSA-C, the olivopontocerebellar pathway is the central focus, grossly portraying a decreased cerebellar size, greatly reduced pons size, blurring of the inferior olive and extensive pallor of white matter within the cerebellum.<sup>201</sup> MSA-P, in contrast, targets the striatonigral pathway. This leads to pallor of the substantia nigra and locus coeruleus, extensive darkening and atrophy of the putamen, yet grossly normal brainstem and cerebellar regions.<sup>201</sup>

### 1.5.2 Understanding MSA etiology

As with many diseases, a sensible route to unraveling MSA is to try to identify and understand the events that increase risk for MSA, and in doing so provide tools with which to model and study the pathogenic process. As with similar diseases, there are two broad areas of risk factor investigation: those of environmental and genetic origin.

Relatives of MSA patients have had significantly more clinical symptoms than did controls; this, along with other work has been used as evidence to suggest a genetic or shared lifestyle etiology component for MSA. It is also noteworthy that the frequency of MSA subtypes varies considerably among distinct ethnic groups: in the British population, MSA-P accounts for an estimated 34% of MSA cases, with MSA-C attributing only 17% and the remaining 49% considered a hybrid of equally severe cerebellar and parkinsonism pathology. On the other hand, MSA-P in the Japanese population is much rarer (17%), while MSA-C is the predominant single subtype, accounting for 40% of all MSA cases, and 42% representing the remaining hybrids.<sup>233</sup> Again, while this cannot be attributed to a genetic, environmental, or lifestyle influence, such variation suggests that there are likely discrete factors that influence this disease.

### 1.5.3 Preliminary association studies

A preliminary investigation of MSA and occupational risk factors suggested that MSA patients had significantly more exposures to a variety of hazardous substances including: plastic monomers and additives, organic solvents, pesticides and metal dusts and fumes.<sup>211</sup> Resonating with PD, occupational farming has been suggested to be a risk factor for MSA, while a history of smoking is associated with a decreased risk for both PD and MSA.<sup>212,234</sup> The role of cholesterol in MSA has also been studied, perhaps in part

because cholesterol has been suggested to interact with alpha-synuclein *in vitro*, potentially altering its conformation and degree of aggregation.<sup>235</sup> One investigation looked at the association between the risk of MSA and serum cholesterol levels, revealing that decreased levels of high density lipoprotein cholesterol (HDL-C) and total cholesterol may be associated with an increased risk of developing MSA, but not duration or severity of disease.<sup>236</sup>

The notion that many disorders are complex diseases embodies the hypothesis that diseases can occur as a result of a complex interaction of genetic, environmental, and lifestyle factors. This concept has been a widely accepted belief that has been applied to many other late onset neurodegenerative diseases. Evidence suggests in diseases such as AD and PD multiple genetic risk factors exist that individually exert small and moderate effects. Though unknown, it is likely that MSA will possess a similar etiologic architecture to these disorders; hence, we should not be looking for either an environmental or genetic cause, but rather accept that the two may coexist as contributors to MSA pathogenesis.

Unfortunately, the relative rarity of MSA and challenge in executing prospective epidemiological studies means that investigation of a potential role for environmental or lifestyle factors in this disease is relatively sparse, and to date no equivocal risk factor has been identified. Further, the environment is an intrinsically difficult entity to study. Upon taking into consideration that different exposures are likely to have unique effects depending on dosage, duration, and timing of exposure, the environment is truly infinite. Conversely, while the genome is certainly large and complex, genetics as a field is adept at using modern methods to elucidate the genetic basis of complex diseases. Therefore, as



relatively minimal information has been revealed within the context of MSA etiology, we propose to address the possible underlying role of genetics.

#### **1.5.4 The genetics of MSA**

##### ***1.5.4.1 Mendelian inheritance***

While reports of possible familial cases of MSA are extremely rare, they have the potential to be very valuable, as unraveling the genetic causes of rare familial forms of disease has provided key insight into several common neurodegenerative diseases like PD. A small number of family based studies reveal kindreds with what appears to be MSA, inherited in an apparent autosomal dominant or recessive inheritance manner.<sup>204,237</sup> While these families are likely to facilitate our understanding of the genetic basis of MSA, to date family-based gene discovery efforts have been few, and thus far not entirely successful in MSA.

With a history of a few genetic studies performed, MSA is currently classified as a sporadic disease; while a few familial studies have argued for an underlying genetic component of MSA, none have been pathologically confirmed among multiple family members.<sup>204,237</sup> In a German family, probable MSA has been reported in a mother and daughter, suggested to be inherited in an autosomal dominant fashion. Despite their similar clinical presentations, the age of onset was greater than 20 years apart (68 for mother, 46 for daughter). While known *SCA* mutations were not identified in either individual, investigators proposed a possible anticipation effect of an unidentified trinucleotide repeat disorder to explain probable MSA phenotypes.<sup>204</sup>

In a multiplex Japanese family consisting of 4 nuclear families, one with a confirmed consanguineous marriage, definitive MSA was reported in one individual while 5 members were diagnosed with probable MSA and 2 with possible MSA. Given the rare estimated prevalence (3-4 per 100,000 among adults 50-99 years of age) of MSA among the general population, the probability of occurrence in two siblings (1 definite MSA, 1 probable MSA) within the same family, by chance, is approximately  $6 \times 10^{-5}$ , making this highly improbable (although not impossible) to occur by chance. Moreover, studies have demonstrated that in the relatives of MSA patients, there is an elevated prevalence of other neurodegenerative diseases.<sup>238</sup> Resonating with the former German study, all hereditary ataxias were excluded and none of the family members harbored any mutations in *SNCA*. While a pattern of autosomal recessive inheritance was proposed, the inability to definitively diagnose more than one affected individual with MSA within each family suggests that preliminary evidence of an underlying genetic etiology cannot be confirmed.<sup>237</sup>

#### **1.5.4.2 *COQ2* mutations**

Perhaps the most progress has been made in this regard with the relatively recent publication by Tsuji et. al, which suggested that rare variants of *COQ2*, the gene encoding coenzyme Q2 4-hydroxybenzoate polyprenyltransferase, play a role in both familial and sporadic MSA. Interestingly, members of a consanguineous Japanese family with MSA-P were reported to be homozygous for *COQ2* variants, p.M78V and p.V343A.<sup>239</sup> The latter variant, p.V343A, which is a common variant within the Japanese population, manifested a significant association with sporadic MSA cases in comparison to controls.<sup>239</sup> Finally, a yeast complementation assay was performed to demonstrate that

p.V343A variants, in addition to other unique variants in *COQ2*, are correlated with dysfunction of *COQ2*.<sup>239</sup> As an antioxidant that prevents free radical damage and mitochondrial oxidative stress, *COQ2* is an intriguing candidate gene to investigate, as it directly parallels our current conceptualization of neuropathology: neuroinflammation induced neurotoxicity and resulting neurodegeneration.<sup>239</sup>

In response to this interesting work, several other groups have attempted independent replication, all with very limited success. Primarily, other factions have clarified that Tsuji et. al used the shortest isoform encoding the smallest protein of *COQ2*, which consequently affects the location of the specified homozygous mutations and does not cover a common nonsense variant at the initial sequence of the first exon.<sup>94,240,241</sup> Upon sequencing *COQ2* in a large Korean cohort, the p.V343A mutation, now designated by its location in the largest isoform, p.V393A, did not portray any association with MSA cases.<sup>242</sup> Furthermore, the study of a large European cohort of clinically diagnosed MSA patients by candidate variant investigation found this same mutation in one case and one control, thus rejecting a potential association between this homozygous variant (p.V393A) and MSA.<sup>243</sup> Finally, Schottlaender et. al used gene sequencing to analyze the most extensive cohort of European pathologically confirmed MSA cases, who found unique *COQ2* variants with a higher frequency in controls than cases, and the absence of p.V393A in both cases and controls.<sup>244</sup> In response, Tsuji et. al have acknowledged these more recent findings and emphasize a more cautious approach to interpretation of their original results.<sup>239</sup> Within the two years, it has been hypothesized that variants in *COQ2*, which can inhibit normal gene function of coenzyme Q10, may prevent oligodendrocyte's ability to maintain lipid-laden myelin sheath, resulting in

increased oligodendrocyte apoptosis and an elevated risk of MSA.<sup>245</sup> While this is certainly interesting work, independent replication is necessary to firmly establish any etiological link of *COQ2* and MSA.

#### ***1.5.4.3 Genes encoding proteins involved in oxidative stress***

It has been suggested that several genes that play a role in oxidative stress, inflammation and mitochondrial dysfunction may exhibit rare variants that increase genetic susceptibility towards the development of MSA.<sup>225</sup> In particular, studies have revealed positive associations between cytokine gene polymorphisms and MSA genetic vulnerability.<sup>68</sup> As cytokines are central players in immunity and inflammation, such findings are consistent with MSA as a neuroinflammatory process. One study investigated eight distinct candidate genes involved in oxidative stress. The data suggested that *SLC1A4*, *SQSTM1*, and *EIF4EBP1* harbored a significant association with MSA, though follow-up investigations are necessary for validation.<sup>246</sup> In addition to cytokines, many chemokines and inflammatory markers are produced upon microglial activation, inducing a neuroinflammatory response.<sup>247</sup> In particular, variants found in *IL-1a*, *IL-1B*, *IL-8* and *ICAM-1* genes have all demonstrated an association with MSA.<sup>248–251</sup> Likewise, a polymorphic region within the tumor necrosis factor (*TNF*) gene, as well as a variant within alpha-1-antichymotrypsin gene, also delineated an association with MSA.<sup>252,253</sup> Once again, because these findings have not yet been convincingly replicated, they should be interpreted with caution.

#### 1.5.4.4 *PRNP*

Interestingly, Shibao et. al reported a case with a patient presenting with both MSA and Creutzfeldt-Jakob disease (CJD). While these two diseases overlap with regard to certain histopathological features, including the atypical abundance of alpha-synuclein proteins within the central nervous system, they had not previously been known to co-exist in a single individual.<sup>240,254</sup> Normal prion protein demonstrates resistance to oxidative stress, but becomes increasingly vulnerable upon conversion to the infectious, pathological isoform. Given the shared histopathology of MSA and prion disease, Shibao et. al hypothesized that the abnormal prion protein may enhance sensitivity towards oxidative stress and consequently contribute to MSA pathogenesis.<sup>254</sup> While homozygosity of the p.M129V allele of prion protein (encoded by *PRNP*) is a known risk factor for CJD, the patient did not harbor any mutations in *PRNP*, but the proband was homozygous MM for the p.M129V allele. To determine if an association indeed exists between MSA and the p.M129V genotype, a case-control study was performed. Results revealed no significant difference in the genotype frequencies between MSA cases and controls, but an elevated prevalence of homozygosity (MM or VV) and younger onset of disease in MSA cases in comparison to PD cases.<sup>254</sup> While this is promising, the absence of abnormal prion proteins within GCIs of pathologically confirmed MSA cases is not trivial, casting doubt on the previous association.<sup>241</sup> Therefore, in order to further elucidate the inflammatory etiology underlying MSA pathophysiology and a potential association with CJD, additional studies to seek out (new and confirm previous) inflammatory marker associations are essential.<sup>225</sup>

#### 1.5.4.5 *SNCA*

The remarkable discoveries of gene mutations in *SNCA* encoding alpha-synuclein have provided key insight into the genetic architecture, pathology, and etiopathogenesis of the most common synucleinopathy, PD.<sup>89,94,97</sup> While Lewy bodies are the hallmark neuropathological findings in PD, they can be identified in approximately 10% of MSA cases. Likewise, mutation(s) of genes classically linked to PD, such as a p.G51D *SNCA* mutation, can also lead to MSA pathology. For example, a recent study of a British patient with autosomal dominant young-onset PD possessing a p.G51D *SNCA* mutation revealed strikingly similar neuropathological and cellular features to a typical MSA case.<sup>208</sup> Although this patient was deemed levodopa responsive, the autopsy exhibited a very high prevalence of GCI-like pathology within the cerebellar white matter, pontine base, and white matter underlying the motor cortex.<sup>208</sup> Further, this case demonstrated positive immunoreactivity for alphaB-crystallin, a GCI-marker; hence, this provides additional evidence for a common pathogenic mechanism behind MSA and PD.<sup>208</sup> In addition to missense (point) mutations, whole gene duplications and triplications of *SNCA* can cause a progressive synucleinopathy through gene dosage elevated expression. Specifically, the *SNCA* gene has been shown to be duplicated or even triplicated in forms of early onset PD, manifesting a Mendelian form of inheritance.<sup>94</sup> Upon studying the neuropathology of affected family members harboring a *SNCA* triplication, the presence of GCIs suggests plausible MSA histopathology.<sup>94</sup> Despite finding GCI-like inclusions in a few cases of PD due to a *SNCA* triplication, studies performing *SNCA* sequencing, gene dosage effects, haplotype tagging and microsatellite analysis of MSA have been unsuccessful in disclosing any disease causing mutations.<sup>97,255–258</sup> Furthermore, studies

scrutinizing gene expression have failed to detect any changes in transcription of *SNCA* among confirmed MSA cases.<sup>233,259–261</sup>

Although no coding mutations in *SNCA* have been identified, a focused genotyping study of MSA revealed a significant association between particular SNPs within the *SNCA* locus and an increased risk of MSA among Caucasians.<sup>206</sup> Follow-up studies initially confirmed these SNP associations, with the most significant located in the MSA-C subtype.<sup>74,262</sup> Residing in the *SNCA* locus, two identified SNPs (rs3822066 and rs11931074), are presumed to be confined within a single haplotype block.<sup>262</sup> This block, extending from intron 4 to the 3' untranslated region (UTR) of the *SNCA* gene, is believed to be in strong LD with the *SNCA* gene.<sup>225</sup> Furthermore, these results have been found in a different cohort of pathologically confirmed MSA cases, garnering further support of an association between MSA and this particular *SNCA* locus.<sup>74</sup> Intriguingly, PD has demonstrated a significant association for this very same haplotype block.<sup>37,263</sup> While this suggests a shared genetic etiology behind PD and MSA, investigations of an association between MSA cases and the risk variants located within this haplotype block have been elusive; indeed, Yun et al. observed an identical allelic frequency of Caucasian risk variants between MSA cases and controls among the Korean population.<sup>225,264</sup> Such studies emphasize the necessity for independent replication across diverse populations, as the inter-population heterogeneity adds an additional layer of complexity to objectively interpret the results of several association studies. In a similar fashion, intra-population heterogeneity has also been shown to be an important consideration: two SNPs in *SNCA*, rs2736990 and rs356220, which have demonstrated to be risk alleles for PD in a Chinese population, failed to manifest any association with either MSA or amyotrophic lateral

sclerosis (ALS) in that same Chinese population.<sup>265</sup> Hence, by performing association studies among several potentially related yet clinically distinct neurodegenerative disorders within a single genetically homogenous population, intra-population heterogeneity may provide insight into the degree of overlap of pathological mechanisms underlying such disorders. Thus, while *SNCA* loci association studies remain intriguing, replication among and within distinct ethnic groups, in conjunction with whole-genome analysis, will be required to confirm or reject these alleged associations.

#### **1.5.4.6 Other PD linked genes**

In addition to *SNCA*, several studies have investigated the frequency of other known PD risk genes and variants among MSA cases. A SNP (rs1572931) within a *RAS* oncogene family-like-1 (*RAB7L1*) promoter region has been demonstrated to be protective in certain populations (Ashkenazi Jews, Chinese) against PD while there has been no association detected, in either MAF or genotype frequency, with MSA for either population.<sup>266</sup>

Several studies have also scrutinized *MAPT*, encoding the protein tau, for variability that may impart risk for MSA. These studies have been inconsistent in their conclusions, with some reporting an association between the H1 haplotype of the *MAPT* locus with both MSA and PD,<sup>37,267</sup> and others reporting an absence of significant associations between MSA and *MAPT* sub-haplotype variants, confounding our current picture.<sup>206,256</sup>

Encoding glucocerebrosidase, *GBA* can harbor mutations that cause the autosomal recessive lysosomal storage disorder Gaucher's disease. Carrying a single *GBA* mutation, while not sufficient to cause Gaucher's disease, is a significant risk factor for PD,



increasing the risk for this disease approximately 5 fold and the risk for DLB at a similar amount.<sup>176</sup> While MSA and GBA-PD portray several overlapping clinical features, screening for the PD associated *GBA* mutation among MSA patients has yet to uncover an association thus far.<sup>268,269</sup>

Mutations in Leucine-rich kinase 2 gene (*LRRK2*), encoding dardarin, have been shown to account for about 3-10% of cases of familial PD and 1-8% of sporadic PD cases.<sup>89</sup> Further, histopathological reports of brains expressing the *LRRK2* mutation also observed overlapping features of MSA neuropathology.<sup>89</sup> While initial association studies between *LRRK2* mutations and MSA have all been negative,<sup>270,271</sup> recent collaborative investigations described a significant association between pathologically confirmed MSA cases and *LRRK2* variants with a protective effect.<sup>272</sup>

As the most prevalent cause of autosomal-recessive early onset PD, mutations in Parkin and PTEN-induced putative kinase 1 (*PINK1*) have been investigated among a pathologically confirmed MSA cohort.<sup>273,274</sup> Results reported the absence of pathogenic homozygous mutations in all MSA cases; while some harbored heterozygous variants, this was not considered a statistically significant association.<sup>274</sup>

Several other genes, including alcohol-dehydrogenase genes, *ADH1C* and *ADH7*, as well as ubiquitin C-terminal hydrolase-1, *UCHL-1*, have been suggested to demonstrate an association with PD.<sup>275–277</sup> After scrutinizing these genes in MSA cohorts, findings have yet to reveal an association for *ADH7* and *UCHL-1* in MSA patients.<sup>278,279</sup> Notably the association at these two genes with PD still remains questionable.

Pathogenic expansion of the hexanucleotide repeat within *C9ORF72* is the most common genetic cause of both ALS and frontotemporal dementia (FTD).<sup>23</sup> Interestingly,

a case study has recently revealed the coexistence of ALS and MSA in a single family.<sup>23,280</sup> While pathological evaluation awaits confirmation of a definite MSA diagnosis, the patient presented a hot cross bun sign on brain MRI. Further, she exhibited ataxia, parkinsonism, autonomic dysfunction and rapid progression, which are all consistent with her diagnosis of possible MSA, while genetic testing of the spinocerebellar ataxias (SCA) was negative. However, Schottlaender et. al and Scholz et. al were both unable to find this mutation among their respective MSA cohorts, suggesting that an association between *C9ORF72* and MSA cannot be validated until MSA is pathologically proven.<sup>281–283</sup> Thus, such a case provides insight into a potentially overlapping genetic etiology between MSA and ALS, despite their unique classical presentation of symptoms.

#### ***1.5.4.7 Copy number changes***

CNVs are structural variants within the human genome which strictly encompass deletion or multiplication of genomic segments that may or may not contain genes.<sup>1</sup> However, copy neutral rearrangements are also part of the CNV family, where a particular segment of genomic DNA is not lost or copied, but rather present in a different position, or orientation within the genome.<sup>1</sup>

As mentioned above, copy number mutation at the *SNCA* locus is already linked to MSA through the presence of GCI pathology in carriers.<sup>94</sup> In part because of technologies that now make discovery and typing of CNVs feasible, there has been heightened interest in the role such structural genomic alterations may play in the disease process.<sup>1</sup> Surprisingly, given that assessment of CNVs remains quite challenging and specialized, MSA has been studied in this regard, although it should be noted the studies

performed thus far are modest in size.<sup>284,285</sup> One investigation performed whole-genome CNV analysis in a 32-person Japanese MSA cohort, as well as a set of monozygotic twins discordant for MSA clinical diagnosis.<sup>284</sup> Analysis described copy number loss of the Src homology 2 domain containing transforming protein 2 (*SHC2*) among the single twin with MSA, as well as 20 of the other MSA patients, while not found in controls.<sup>284285284283284276</sup> As CNVs are known to induce genomic instability and can contribute to unequal crossing over or end-joining events during meiosis, the results suggest that this CNV-rich subtelomeric site may be vulnerable to insertion, deletion or duplication events.<sup>286</sup> Furthermore, CNVs in genes are known to have several potentially deleterious effects, including modified expression in a cis or trans fashion, and the formation of unstable mRNA and protein products, possibly responsible for pathophysiology.<sup>287</sup> Since Shc proteins play a role in neuronal cell development, acting as molecular switches for proliferation and differentiation, the potential for pathophysiology is not unlikely.<sup>284</sup>

Given the discordance among monozygotic twins certain environmental factors may be critical for turning on and off genes, thereby modulating genetic expression and possibly inducing MSA pathophysiology.<sup>288</sup> Ferguson et. al was unable to find CNVs in the *SHC2* gene among a non-Japanese MSA cohort in a follow-up study.<sup>285</sup> Thus, while *SHC2* CNV analysis requires independent replication in a larger Japanese cohort and among diverse populations, the results from Sasaki et. al are promising.

While progress is being made towards elucidating the genetic basis of this disease, more needs to be done. This is particularly challenging in a disease such as MSA, not only because funding is limited, but also because this is a rare disease, and

many of the state-of-the-art methods require large numbers to yield sufficient statistical power. Nonetheless, the existing opportunities for the genetic dissection of complex disease are markedly better than a decade ago, and it is essential that we attempt to push genetic progress in MSA.<sup>201</sup>

### **1.5.5 Proposed mechanisms of MSA pathogenesis**

Although little is known about the genetic etiology of MSA there has been some work focused on understanding the molecular pathogenesis of this disease.

Predominantly, this has been derivative work ongoing in PD rather than based on unique molecular aspects of MSA.<sup>202</sup>

#### ***1.5.5.1 Role of Neurotoxicity and Oxidative Stress***

Given that microglial activation is associated with neuronal loss, the initiation of extensive microglial over-activation in olivopontocerebellar and striatonigral regions of the brain in MSA is intriguing.<sup>202,289</sup> One study observed microglial transition into a state of over-activation upon exposure to environmental toxins and endogenous proteins.<sup>290</sup> This microglial excitability, specifically triggered by pattern recognition receptor transduction mechanisms, initiates a release of reactive oxygen species (ROS), well-known culprits of inducing neurotoxic states.<sup>290</sup>

Investigators induced alpha-synuclein overexpression (with a PLP promoter) in conjunction with exposure to toxin 3-Nitropropionic acid (3-NP) in a transgenic mouse model. Histopathological analysis revealed GCI-like inclusions with a substantial loss of neurons in regions primarily targeted by MSA pathology: olivopontocerebellar and striatonigral systems.<sup>289</sup> Phenotypically, there was a depreciation of motor and cerebellar function. Interestingly, elevated levels of inducible nitric oxide synthase (iNOS), which

plays a role in immunity and free radical propagation, was reported in the SNPC.<sup>289</sup>

Further, a direct correlation was observed between increased iNOS levels with both the disappearance of striatonigral dopaminergic neurons and a rise in microglial activation, particularly in the SNPC.<sup>289</sup>

These findings provide key insights into our understanding of MSA pathogenesis. Principally, they suggest an increased susceptibility of this region to oxidative stress, which may serve as an impetus for neuroinflammation.<sup>289</sup> Based upon this notion, anti-neuroinflammatory agents have been tested in transgenic mice. Despite its anti-neuroinflammatory properties being somewhat elusive, long-term minocycline treatment was administered in the transgenic mice. Consequently, microglial activation was inhibited in the SNPC, protecting dopaminergic neurons in this area.<sup>289</sup> While the mechanism of action is unknown, potentially neuroprotective agents warrant further investigation, as it appears that oligodendroglial overexpression of alpha-synuclein in GCIs and oxidative stressors are definite culprits in this devastating neurodegenerative disease.

As a constituent of the lipid component of the cell membrane, Docosahexaenoic acid (DHA) can increase cell sensitivity to oxidative stress.<sup>291</sup> When elevated levels of DHA are present within the cell membrane, heat shock protein expression increases, which is known to rise under conditions of oxidative stress.<sup>291</sup> Regarding MSA pathology, oligodendroglial cells with heightened DHA levels, expressing alpha-synuclein, are increasingly sensitive to oxidative stress. Moreover, this rise in oxidative stress sensitivity actually makes alpha-synuclein more insoluble, forming fibrillary inclusion bodies like those in MSA.<sup>292</sup> Such aggregate formation is simultaneously

enhanced through a rise in phosphorylation of alpha-synuclein at serine-129, which resonates with classic MSA pathology.<sup>292</sup>

Further investigations of oxidative stress have studied myeloperoxidase, a crucial enzyme that plays a role in phagocytosis associated cell production of ROS.<sup>293</sup> Since it exists in both human and mouse brains, and myeloperoxidase-containing macrophages and microglia have been reported in the CNS among other neurodegenerative diseases including PD, myeloperoxidase manipulation serves as a useful enzymatic tool to elucidate the role of neuroinflammation and oxidative stress in MSA.<sup>293</sup> Numerous experiments have observed that neuroinflammation is a “prominent pathological finding” in MSA, which is a clear facilitator of oxidative stress.<sup>247,293</sup> While the current mechanism inducing neuroinflammation in MSA is uncertain, it is hypothesized that potentially rare variants of genes associated with inflammation may enhance susceptibility to such neuroinflammation.<sup>247,293</sup> Primarily, it has previously been demonstrated that myeloperoxidase is involved in the neuroinflammation and neurotoxicity of MPTP induced PD, suggesting a potentially neuroprotective role of myeloperoxidase inhibition. In a transgenic mouse model, inhibition of myeloperoxidase has several profound effects. Primarily, it has the ability to protect neurons vulnerable to oxidative stress in the SNPC, cerebellar cortex, striatum, pontine nuclei and inferior olives.<sup>293</sup> Secondly, reverberating with the results of minocycline administration, myeloperoxidase inhibition decreases the amount of microglial activation, though notably does not influence astrogliosis.<sup>293</sup> Thirdly, it results in a reduction of intracellularly located alpha-synuclein aggregates, suggesting a potential therapeutic role of mitigating inflammation and oxidative stress. This decrease of alpha-synuclein aggregates occurs in

a dose-dependent manner, with higher doses of a myeloperoxidase inhibitor corresponding to larger declines in alpha-synuclein positive GCIs and elevated neuronal survival in the SNPC and striatum.<sup>293</sup> From a phenotypic perspective, the reduction of motor dysfunction suggests a “partial reversal of oligodendroglial alpha-synuclein nitration and aggregation.”<sup>293</sup>

A thorough analysis of MSA literature illustrates that many groups have garnered evidence to hypothetically explain the accumulation of neurotoxic alpha-synuclein aggregates in GCIs: alpha-synuclein is derived from neurons but spreads to oligodendroglia. In 2012, Kisos et al revealed that in the presence of elevated alpha-synuclein levels, either in the form of soluble oligomers or intracellular alpha-synuclein inclusions in neurons, neuronal secretion is enhanced within rat brains.<sup>294</sup> Specifically, it was demonstrated that rat oligodendroglial cells *in vitro* internalized alpha-synuclein from neuronal secretions in a time, concentration and clathrin-dependent fashion.<sup>294</sup>

Furthermore, Rockenstein and colleagues designed transgenic mice models to study heterozygous progeny. Among the parental mice, one expressed alpha-synuclein under an oligodendroglial-specific myelin-basic promoter and the other parental mouse expressed alpha-synuclein under a neuronal platelet derived growth factor promoter. Studying the compound transgenic mice progeny demonstrated a “robust redistribution” of alpha-synuclein.<sup>295,296</sup> While the exact mechanism of action is unknown, Rockenstein and colleagues hypothesized that a direct “translocation” through the extracellular space occurred via cell-cell interactions, moving alpha-synuclein from neurons to neighboring oligodendrocytes.<sup>295</sup>

Collectively, these data suggest a predilection for alpha-synuclein accumulation in oligodendroglia relative to the neurons in regions of the brain susceptible to MSA, resonating with classic pathophysiological changes seen in the disease.<sup>295,296</sup> Further, in 2014, Reyes et al demonstrated that oligodendrocytes can successfully uptake recombinant alpha-synuclein and internalize it in vivo in mouse cortices.<sup>297</sup> Thus, while evidence for this mechanism is substantial, recent findings may suggest that several mechanisms occur in tandem.

Regarding the transmission of alpha-synuclein in MSA, Asi et al demonstrated in 2014 that alpha-synuclein mRNA is expressed in oligodendrocytes among MSA post-mortem brain tissue.<sup>137,138</sup> While we know alpha-synuclein is transcribed and translated in neurons, the possibility of glial cells transcribing alpha-synuclein is intriguing, as it suggests that some of the alpha-synuclein aggregates in oligodendrocytes may indeed originate from those cells, or may even be transmitted to neurons to form NNIs and NCIs. Taken together, the newly recognized significance of neuronal pathology in MSA (i.e. NCIs) and proof of alpha-synuclein seeding and propagation mechanisms represent important milestones in unraveling MSA pathophysiology and have since been incorporated into our evolving framework of neuroinflammation and neurotoxicity.

#### ***1.5.5.2 Role of ubiquitin-proteasome system***

Along with the mechanisms of neuroinflammation and neurotoxicity, the role of protein turnover through the ubiquitin-proteasome system (UPS) and its association with MSA pathophysiology has garnered interest within recent years. Prior studies of alpha-synucleinopathies like PD have illuminated the role of UPS dysfunction.<sup>139</sup> The failure of



the UPS in the substantia nigra correlates with the presence of Lewy bodies seen in PD.<sup>139</sup>

Degradation of alpha-synuclein occurs by either one of two cellular mechanisms: autophagy or proteasomal machinery.<sup>140</sup> The former entails a lysosomal pathway forming autophagosomes, which utilize autophagosomal protein markers, *LC3* and a ubiquitin binding protein, *p62*, to induce entry of polyubiquitinated proteins, targeted for cellular destruction, inside the autophagosomes.<sup>140</sup> The study of pathways used for oligodendroglial acquisition of alpha-synuclein accumulations in seven MSA cases have detected *LC3*-positive vesicles demonstrating an association with the alpha-synuclein aggregates located within GCIs. Given that *LC3* is an autophagy lysosomal pathway protein marker, this indicates a potential upregulation of this pathway in MSA pathophysiology.<sup>140</sup> Notably, it was specified that only a subset of the GCIs were *LC3* positive, suggesting that increased activity of the autophagy pathway occurs after alpha-synuclein aggregations have already formed.<sup>140</sup> Further, there is evidence of “genuine cross-talk” between the autophagy and UPS pathways, which may indicate a simultaneous downregulation of the proteasomal pathway in MSA pathogenesis.<sup>140–142</sup> While a mechanism for such communication is under scrutiny, studies have revealed that a reduction in UPS pathway activity leads to elevated stress of the endoplasmic reticulum due to an accumulation of aggregated ubiquitinated proteins. Consequently, this unfolded protein response (UPR) forms a pathway between the endoplasmic reticulum and cell nucleus whereby transcriptional upregulation for genes that activate the autophagy lysosomal pathway.<sup>141</sup> Thus, while several neurodegenerative disorders have been associated with a decrease in UPS pathway activity, this may induce a corresponding rise

in the autophagy pathway.<sup>140,143</sup> With a potentially interdependent system between autophagy and proteasomal pathways, it is believed that while compensatory changes can be made in an effort to maintain a necessary protein degradation balance, perturbations in either system can have pronounced adverse effects.<sup>140,143</sup>

In addition to testing the role of UPS dysfunction and MSA pathogenesis *in vitro*, transgenic mouse models have been designed to enhance our understanding. By using transgenic mice expressing human alpha-synuclein, one investigation confirmed that the UPS is the primary degradation pathway for alpha-synuclein under normal conditions *in vivo*.<sup>142</sup> However, an abundance of alpha-synuclein within human alpha-synuclein transgenic mice due to a dysfunctional UPS induced activation of the autophagy lysosomal pathway, presumably as a compensatory mechanism.<sup>142</sup> Further, a well-established pattern of this altered pathway regulation sequence occurred with a greater frequency in aged mice. As a mechanism that fits in an age associated disorder, this may suggest that increased age, in conjunction with an elevated alpha-synuclein burden, is a risk factor for increased proteasomal pathway dysfunction.<sup>142</sup> With consistently increased alpha-synuclein levels, the UPS pathway may be disrupted; Despite the autophagy pathway's compensatory efforts to upregulate protein degradative functions, a vicious cycle ensues, culminating in vast accumulation of alpha-synuclein in GCIs and oligodendroglial cell death.<sup>144</sup>

Further studies in transgenic mice have explored the phenotypic modifications associated with UPS dysfunction. Mice expressing human oligodendroglial alpha-synuclein experienced proteasomal pathway inhibition via induction of systemic proteasome inhibition (PSI).<sup>144</sup> Specifically, PSI activation resulted in motor dysfunction,

which was directly correlated with neurodegeneration in the striatonigral and olivopontocerebellar systems of these transgenic mice. In contrast, mice expressing human oligodendroglial alpha-synuclein but lacking PSI induction manifested an absence of motor deficits and neuronal loss in corresponding regions.<sup>144</sup> Furthermore, systemic application of PSI in transgenic mice resulted in selective neurodegeneration of striatonigral and olivopontocerebellar systems, while all surrounding areas were unaffected, resonating with human MSA's affected regions.<sup>144</sup>

It is evident that PSI treatment in the transgenic mice induced aggregation of human alpha-synuclein located within oligodendroglia, as manifested by GCIs. This may have resulted in myelin degeneration, axonal swelling, and mitochondrial enlargement, a clear sign of mitochondrial stress. Identical to MSA neuropathological findings, such transformations suggest that UPS dysfunction plays a central role in the mechanism of MSA pathogenesis.<sup>144</sup>

### **1.5.6 The dynamic behind key players: neurotoxicity, oxidative stress and the UPS**

To connect several key findings regarding the molecular mechanisms of MSA pathogenesis, it is useful to study the relationship between the UPS and autophagy pathways with oxidative stress. Recent investigation of the ubiquitin homologue, SUMO-1, has identified it within “discrete subdomains” of alpha-synuclein inclusion bodies of MSA brain tissue.<sup>300</sup> Interestingly, in the brain tissue of MSA and PSP cases, a co-localization was reported between a lysosomal subset and SUMO-1. As neurodegenerative diseases both exhibiting cytoplasmic inclusion bodies of alpha-synuclein and tau, respectively, these findings may indicate an association between protein aggregation and SUMO-1 via the lysosomal autophagy pathway.<sup>300</sup> As prior

investigations have strongly suggested a downregulation of UPS and an upregulation of the autophagy lysosomal pathway in MSA pathogenesis, SUMO-1 could play a key role in the pathophysiology.<sup>300</sup>

### **1.5.7 Drug Therapies and targets in MSA**

As MSA and PD are both members of the alpha-synucleinopathy family, it has been suggested that drug discovery for both neurodegenerative diseases should target their overlapping pathophysiology.<sup>301</sup> Specifically, while the MSA-P subtype has been described to exhibit several shared clinical features with PD, it has been deemed more rapidly progressive and fatal.<sup>301</sup> Given the report of a British individual harboring a *SNCA* p.G51D mutation with pathologically confirmed GCIs and LBs, a shared mechanism of disease is indeed plausible.<sup>208</sup>

Using functional imaging with both florodopa and b-CIT single photon emission computerized tomography (SPECT), investigators have been able to track the annual loss of signal among brains lesions in vivo in both MSA-P and PD cases. Notably, the estimated annual loss of brain signal in PD has been suggested to be 5-10%, while the MSA-P progression rates have been reported to be much higher.<sup>301,302</sup> Further, using MRI, the regional atrophy exhibited in patients with MSA-P has been approximated at a 1-2.5% annual decrease, while only 0.3-0.8% for PD, respectively.<sup>301-305</sup> Moreover, using positron emission tomography and amyloid ligand benzoxazole, it has been reported that GCIs in MSA patients can be visualized in vivo, making this an excellent potential drug target.<sup>306</sup>

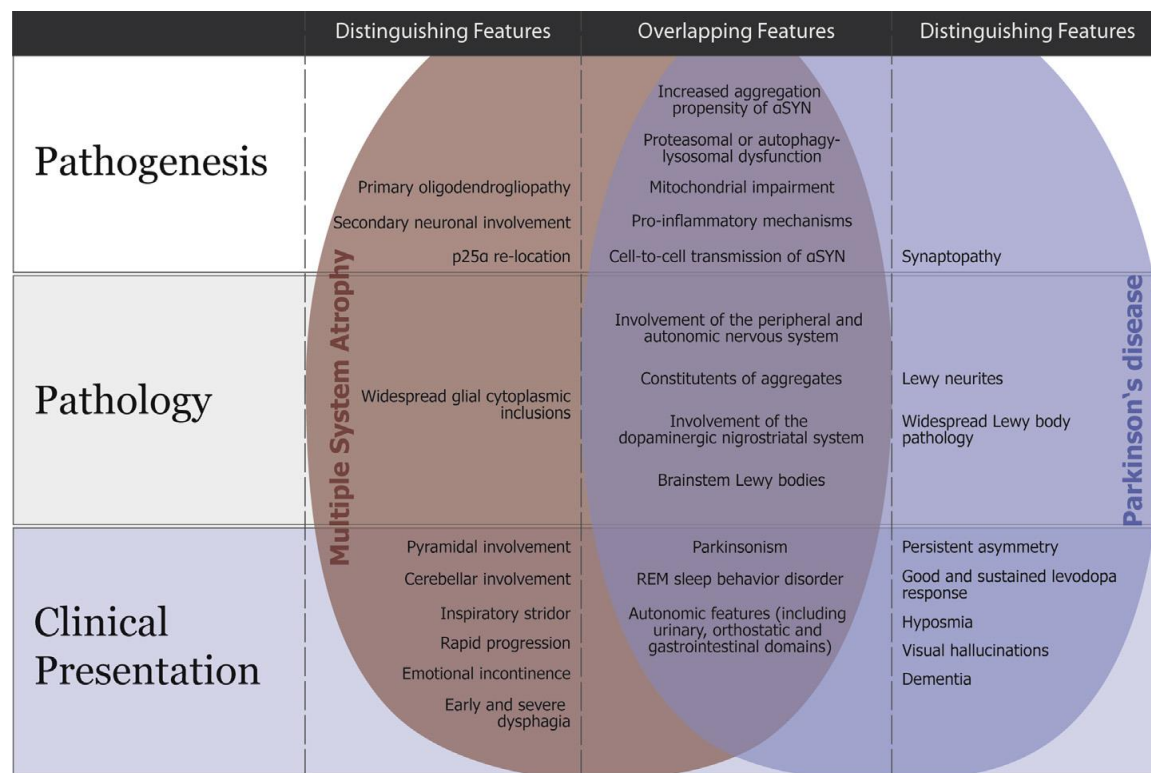
Given MSA-P's ability to reveal pathological progression in an accelerated and quantitative fashion as compared to PD, it has been suggested that taking advantage of

these properties will facilitate a more expedient and steadfast approach to understanding alpha-synuclein pathology.<sup>301</sup> Described as a “MSA proof of concept trial,” the benefits extend beyond both time and cost-efficiency, further reasoning that given the lack of symptomatic treatment for MSA-P, several short-term MSA clinical trials could run in parallel, not confounded by the use of any symptom modifying therapies (i.e. carbidopa-levodopa in PD).

By focusing drug therapy efforts for alpha-synucleinopathies on MSA-P patients, the importance of obtaining accurate diagnoses becomes critical. While the current consensus criteria for possible MSA has been reported to show an estimated 95% positive predictive value between the initial clinic visit and post-mortem MSA diagnosis, the need for a plasma or CSF biomarker is crucial to achieve the greater sensitivity and specificity.<sup>301</sup> In recent months, Mitsui et al. have reported significant differences in plasma CoQ10 levels between MSA patients and controls after adjusting for age, sex and COQ2 genotype.<sup>307</sup> Other studies have also compared plasma CoQ10 levels in PD patients. While a significant difference in CoQ10 plasma levels between MSA and PD patients has yet to be reported in such studies, larger samples are likely required to obtain statistical significance.<sup>308</sup> Finally, recent CSF studies have compared levels of neurofilament light chain and microRNAs between MSA patients, PD patients and controls to determine if either has the potential to serve as a biomarker.<sup>309,310</sup> Preliminary studies have demonstrated statistically significant results for both molecular entities, suggesting yet another avenue to pursue regarding MSA diagnostic accuracy.<sup>309,310</sup>

### 1.5.8 A comparison of PD and MSA

While progress is being made towards finding an extremely sensitive and specific biomarker to differentiate PD and MSA, some of the pathogenic, pathologic and clinical features are notably distinct (Figure 11).



**Figure 11: Shared and distinguishing pathogenic, pathologic and clinical features of MSA-P and PD.**

Regions highlighted in red reflect those unique to MSA-P. Regions highlighted in blue reflect those unique to PD. Regions highlighted in purple reflect those shared by both MSA-P and PD.

(Reproduced by Krismer 2014 et al.)

While the pathogenic and pathologic features continue to be explored through several types of functional imaging, the clinical presentations have several differences that facilitate diagnosis in the prodromal stages of disease. For example, while certain autonomic features are common to both neurodegenerative diseases, specific autonomic symptoms, such as dysphagia, is particularly unique to MSA. Likewise, a very specific

sensorineural phenotype, hyposmia, is specific to PD.<sup>301</sup> While one cannot exclude either of these diseases in the differential diagnosis secondary to the presence or absence of specific clinical phenotypes, using well-defined clinical information as a guide to seeking future testing (i.e. biomarker, imaging) may play an instrumental role in our evolving understanding of both diagnosis and treatment of MSA and PD.

### **1.5.9 How to move forward**

Understanding the disease process is a crucial milestone in the development of etiologic therapies; however, as is illustrated above, so much uncertainty remains regarding the molecular underpinnings of MSA. We believe that a priority in elucidating this disease lies in defining and identifying the genetic architecture. This would not only provide a window into the etiology but will likely be critical for biomarker development and in the early identification of pre-symptomatic patients. To discern the specific types of variants that may be involved, it is important to consider two distinct but not mutually exclusive paradigms: CDCV and CDRV hypotheses.<sup>15</sup> In a complex disease it is reasonable to suggest that there is a synergistic effect among common and rare variants that all contribute to disease risk and development. This theory, described as Pleomorphic risk locus (PRL) hypothesis, accounts for the underlying complexity behind polygenic disorders that can present with extensive phenotypic variability and severity.<sup>15</sup> While GWA studies are ideal for the pursuit of common variants and risk association, NGS via WES is promising for rare variants, as there have been many successful novel variant discoveries among complex neurodegenerative diseases in recent history.<sup>21,113</sup> Furthermore, the former SNCA example in PD illustrates the notion that both common and rare variants present on the same loci can contribute to varying degrees of risk.<sup>15</sup> In

essence, these contributing loci can be called ‘modifiers’ to disease risk, unlike the classic Mendelian monogenic inheritance patterns.<sup>9</sup> Applying the theory of PRL to MSA, we hope to discover the association of common variants following imputation of GWA study data. However, if we find any of significance, this is likely only to comprise a small fraction of MSA risk. Thus, WES followed by targeted resequencing will play a key role in unraveling and validating novel rare variants that influence one’s risk of developing MSA.

Familial studies, SNP and gene association studies have highlighted the role of genetics in MSA from an etiological perspective. Given that MSA is largely unresponsive to levodopa and current treatment is primarily oriented to symptomatic relief, the significance of unraveling the genetics and etiology of MSA is paramount, as there is an urgent need to move toward etiologic based therapies. With very limited insight of genetic mutations or alterations in gene dosage as a cause of MSA, the hunt for novel risk genes, which may be in the form of common variants or rare variants, is the logical nexus for MSA research.<sup>26</sup> Prior investigations have studied the role of potential environmental risk factors, with some reporting that MSA patients have been exposed to environmental insults more than controls.<sup>60</sup> While intriguing, the feasibility of pursuing further study in this domain is challenging; specifically, identification and quantification of the numerous possible toxicant exposures that may contribute to MSA pathogenesis is challenging.<sup>60</sup> Conversely, pursuing genetic risk and causative loci is scientifically tractable. Implementation of next generation methods, including genome wide association (GWA) and second generation sequencing, provide the ability to obtain valuable data and inform clinical diagnosis.<sup>22</sup>



Success in any modern genetic investigation of MSA will require extensive scientific collaboration, regardless of approach. MSA is rare enough that no single group can collect sufficient cases on its own; thus, the field will not progress without pooling of clinical resources. In an effort to efficiently and easily share resources, the burden of analysis, and rapidly disseminate results, the formation of an international collaborative framework should be the priority of any entity wishing to pursue research into the genetic basis of MSA. Upon acknowledging these substantial challenges, we would predict that clinical progress of MSA (diagnosis, treatment) will be much delayed until we make advances towards our genetic understanding of this disease.

## 2 Estimating the heritable component of MSA

*Statement of contribution:* Genotyping was performed by the Genomic Technologies Group of the Laboratory of Neurogenetics. The first phase of the GWA study was performed by Anna Sailer in collaboration with the Statistical Genetics Group of the Laboratory of Neurogenetics. I performed data quality control, genotype imputation analysis, and the subsequent execution of the heritability analysis using GCTA with T.R Price from the Statistical Genetics Group. I also contributed to the final draft of the first MSA GWA study, currently under review in *Neurology*.

### 2.1 Introduction

Within the last decade, there has been a substantial increase in the number of GWA studies investigating many traits, including disease.<sup>311</sup> These case-control studies are pursued with a primary goal of determining which variants associate with a particular phenotype. This approach is favorable towards the identification of common genetic risk factors for disease phenotypes in a specific population. Previous investigation of several neurodegenerative diseases demonstrates the power of GWA studies and its ability to identify key risk loci.<sup>169,187,312–314</sup> Despite the fact that several loci have been discovered in complex diseases such as PD, one must recognize that the identified loci only explain a relatively small proportion of the total heritable component of disease. While the known GWA loci only account for 3-12% of the burden of PD, current conservative estimates of the heritable component of this disease are ~30%. It is evident that increasing our understanding of the known and unknown heritable components of disease can be highly

informative in the research community, particularly for ascertaining the value of searching for additional genetic risk and corresponding genomic locations.

Using 1,030 MSA samples, a recent GWA study for MSA risk loci assessed more than 5 million SNPs tagged to common genetic variants (Sailer et al, under review 2016). After quality control measures, the results included 918 MSA cases and 3884 controls but failed to detect any genome-wide significant associations between tagged SNPs and MSA risk. This finding suggests that MSA etiology cannot be easily explained by common SNPs with moderate or large effects, though one must acknowledge the limited sample size and power in this study. In PD, for instance, GWA studies required more than 1400 samples to identify significant associations.<sup>37</sup> Hence, as we recognize the possibility that we may be underpowered, this result does not preclude the role of common variability in MSA. Thus, variants conferring marginal effects, typical of those observed for GWA in complex disease, may impose risk towards the development of MSA. The opportunity to look beyond the identification of individual risk loci, toward an estimate of the role and extent of common variability in risk for MSA is achievable using this genotyping data set.

To estimate the total heritability of MSA from common genetic variants ( $MAF > 0.01$ ), we utilized an approach using Genome-wide complex trait analysis (GCTA).<sup>48</sup> By defining heritability as the phenotypic variation attributable to total genetic variation in all assessed loci, we could estimate the total genetic variation by creating a genetic relatedness matrix (GRM) in GCTA. In essence, the GRM estimates overall genetic differences in each subject; hence, if cases are more genetically similar to one another than they are to controls, we can quantify this higher relative similarity and use it to

estimate the total heritability of the disease phenotype. Notably, substantially large sample sizes in unrelated populations are required for requisite statistical power when using GCTA. This ultimately allows one to measure the overall polygenic additive inheritance by incorporating putative causal variants that are in complete linkage disequilibrium (LD) with common SNPs but have minimal effect size.<sup>203</sup> Given that GCTA often incorporates imputed data from genotyping microarrays, it typically only assesses the effect of putative causal variants in LD with all common SNPs on the genotyping platform.<sup>48,315</sup> With a principal goal of guiding future genetic research in MSA, we estimated the total heritability of MSA with GCTA.

## **2.2 Materials and methods**

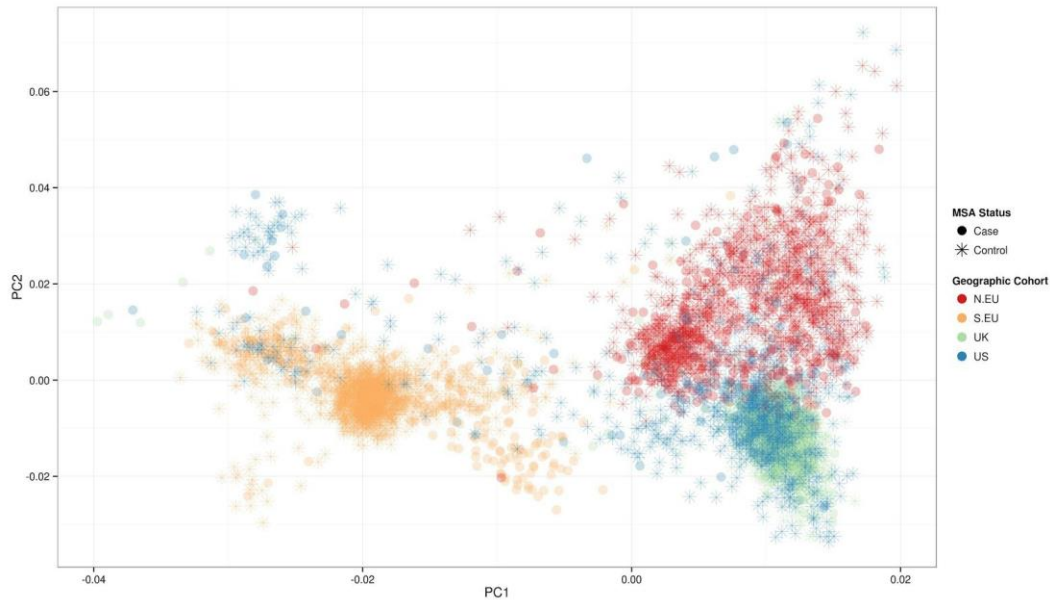
### **2.2.1 Subjects**

A total of 1030 MSA DNA samples were obtained from 4 geographic regions: United Kingdom, United States, Southern Europe and Northern Europe. Southern European nations consisted of Italy, Spain and Portugal; Northern European nations comprised Germany, Austria, Denmark and the Netherlands. Among this cohort, 699 MSA samples received clinical diagnoses from movement disorders specialists and 331 MSA samples were pathologically confirmed by neuropathologists. A total of 3884 neurologically normal controls were obtained from the following 4 nations: United Kingdom (n = 936 samples), Germany (n = 944 samples), United States (n = 794 samples), and Italy (n = 1,190 samples). Since samples were derived from different geographic regions across Europe and the United States, we matched cases with regional controls. Thus, UK MSA cases were matched with UK controls; Northern European

MSA cases from Germany, Netherlands, Austria, and Denmark were matched with German controls; Southern European cases from Italy, Spain and Portugal were matched with Italian controls; American MSA cases were matched with American controls. All samples included in this study were of self-reported European descent. Written informed consent was obtained by all subjects.

### **2.2.2 Pre-imputation base calling quality control**

All variants with <95% call rate across all samples as well as individuals with <95% total variant call rates were excluded from analysis. Using identity by descent (IBD) analysis in PLINK v1.90, we identified and discarded all samples who were more closely related than 0.125 (first cousins).<sup>316</sup> If individuals were 6 or more standard deviations from the average homozygosity of the sample population they were also eliminated. Finally, we excluded all variants that significantly deviated from Hardy-Weinberg-equilibrium (HWE) ( $p < 10^{-5}$ ) in addition to those with a minor allele frequency (MAF) of <0.01. Though checking for deviations in the population is the main purpose of HWE, it also serves as a subsequent filter to exclude genotype assays with suboptimal performance.<sup>317</sup> Utilizing 50-SNP windows with a variance inflation factor (VIF) of 0.5, we pruned the remaining SNPs for LD in PLINK. Since large genetic differences between case and control populations could be misinterpreted as a genetic variation associated with disease, we eliminated individuals whose principal component value was more than 9 standard deviations from the average of either of the top two principal components of 1K Genomes European Ancestry (Figure 12). Using only our cases, controls and linkage-pruned SNPs that passed quality control (QC) filters, we proceeded with imputation.



**Figure 12: Principal components of MSA samples.**

Population stratification using the top two principal components from genotyped SNPs demonstrates that MSA cases and controls cluster uniformly with respect to their geographic origin.

(Reproduced from Federoff and Price et al 2015).<sup>318</sup>

### 2.2.3 Imputation

The process of imputation infers sample genotype data from a reference haplotype database. Autosomal genotypes were imputed using the November 2012 release of the 1K Genomes haplotype reference by matching the genotypes to common haplotypes.

Next, we used a program called Markov-Chain based haplotyper (MaCH) to estimate subject haplotypes. This allowed us to perform the imputation and assess imputation accuracy and quality by removing SNPs with an R-squared (correlation between expected and observed genotype)  $< 0.30$  and MAF  $< 0.01$ .

Using minimac on default settings to impute the haplotypes, 11,138,628 variants passed this imputation thresholding.<sup>319</sup>

With our newly updated and imputed dataset, we replicated the GWA study from Sailer et al (under review 2016) using mach2dat to assess association in MaCH output, adhering to logistic regression under an additive model and using the top 20 population principal components as covariates.<sup>30</sup>

#### **2.2.4 Genome-wide complex trait analysis**

In order to estimate the variance in phenotype explained by variance in genotype, GCTA uses a REstricted Maximum Likelihood (REML) model. After adjusting for population substructure using the top 20 European Ancestry population principal components, we incorporated GCTA's REML model to estimate the phenotypic variance of MSA. Given the rarity of MSA, this heritability estimate was adjusted for actual population prevalence of MSA (estimated at 0.000046).<sup>320,321</sup> In the first analysis, we first ran GCTA using all samples in a pooled analysis. Subsequently we then divided MSA cases into several sample subsets based on geographic region of origin and whether cases had received pathology-confirmed or clinical diagnoses. Further, we tested each of these groups against controls to estimate total heritability of MSA both preceding and following imputation. Using a random effect models, we ran a meta-analysis of these subgroups to obtain heterogeneity assessments between the cohorts.

#### **2.2.5 Bayesian estimate of PD-derived heritability**

Using false diagnostic rates reported by Osaki et. al 2009, we attempted to estimate the rates of clinical misdiagnoses in our MSA cohort.<sup>322</sup> Incorporating a 6-25%

false positive rate in MSA diagnoses, with 70% of false positives (type I error) as Parkinson's cases, and heritability priors of 0.31 for PD and 0.1 for all other disorders (i.e. PSP, DLB, CBD), we calculated an expected degree of heritability to due misdiagnosis with the following formula:

$$\text{Clinical cases (Clin)} * \text{false positive rate (FPR)} = \text{Misdiagnosed cases (M)}$$

$$0.7M = \text{PD cases (P)}$$

$$0.3M = \text{Other misdiagnoses (O)}$$

$$O + P = M = \text{Clin} * \text{FPR}$$

$$(0.31 * P + 0.1 * O) / \text{Total cases} = \text{MSA Heritability due to misdiagnosis (H}_m \text{)}$$

$$H_m = (0.31(0.7M) + 0.1(0.3M)) / \text{Total}$$

$$\frac{0.247M}{\text{total}} = H_m$$

Which simplifies to:

$$H_m = 0.247 (\text{FPR} * \text{Clin}) / \text{Total}$$

## 2.3 Results

### 2.3.1 Quality Control

Following initial QC filters of genotyped data and the linkage-pruned SNP data sets, we were able to perform all PCA and IBD analyses. As illustrated in **Figure 12**, MSA cases and their respective geographic cohort controls cluster uniformly in a principal component analysis of genotypes. This suggests an insignificant amount of population heterogeneity within each regional cohort. Those passing initial quality control filters included 907 MSA cases, 3,877 controls, and a total of 107,447 SNPs (**Table 3**).



<b>Cohort</b>	<b>Number of subjects</b>	<b>Path confirmed / Clinically Diagnosed</b>
<b>Cases</b>		
United Kingdom	238	141/97
United States	127	108/19
North European	308	28/258
South European	234	14/220
<b>Total</b>	<b>907</b>	<b>291/616</b>
<b>Controls</b>		
United Kingdom	945	----
United States	793	----
North European	944	----
South European	1184	----
<b>Total</b>	<b>3877</b>	

**Table 3: Summary statistics of samples included in GCTA analysis.**

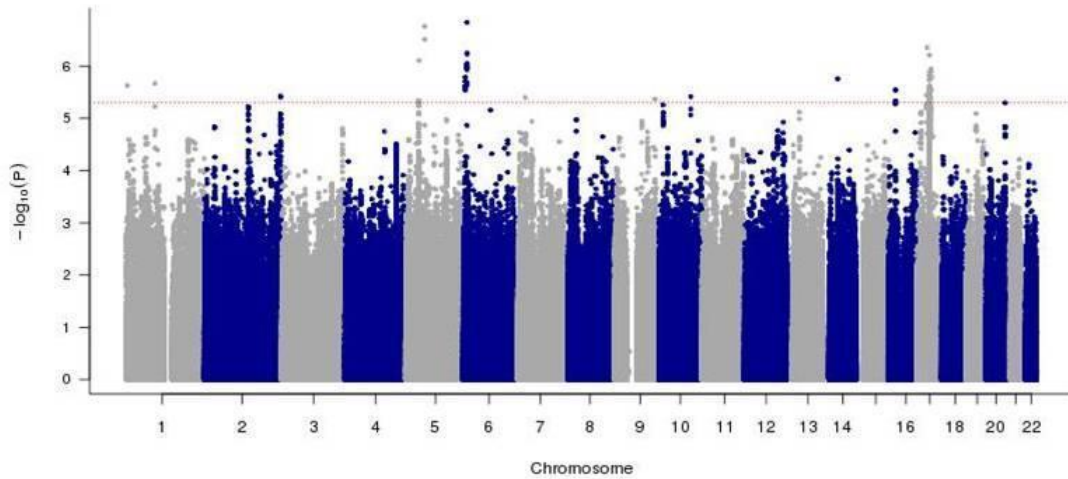
Summary statistics of all samples included within GCTA following stringent quality control analysis. Component numbers of control subjects do not sum to total due to incomplete annotation (i.e. unknown region of origin). Cases not explicitly labeled as pathologically confirmed were assumed to have only a clinical diagnosis.

(Reproduced from Federoff and Price et al 2015).<sup>318</sup>

Next, we performed imputation using 1k Genomes reference haplotypes to ultimately increase statistical power and incorporate several more variants for assessment of total heritability.

### **2.3.2 Post-imputation GWA**

Using a significance value ( $p < 5 \times 10^{-8}$ ), no variants were deemed statistically significant (Figure 13).



**Figure 13: Manhattan plot of post-imputation MSA GWA study.**

P values are log transformed (y-axis) and plotted against chromosomal position (x-axis). The dotted line indicates threshold of potentially interesting SNPs. After Bonferonni correction, none of these SNPs were statistically significant.

(Reproduced from Sailer et al. 2016).<sup>323</sup>

### 2.3.3 Post-imputation candidate gene analysis

Given our suspicion that MSA heritability estimate results may be driven, at least in part, by misdiagnosed PD cases, we replicated the GWAS performed by Sailer et al (under review 2016) with our updated imputation dataset and explored windows  $\pm 20$  kilobases around PD GWA study loci derived from Nalls et al 2014 and from the closest genes associated with mRNA expression differences.<sup>169,324</sup> Despite using an extremely liberal significance cutoff of  $p < 0.05$ , no variants included in this search manifested an association with MSA disease phenotype. However, given our limited sample size, we acknowledge the likelihood of insufficient power to detect such variants. The most

significant variants (labeled by rs number) included in these windows are listed in Table 4.

SNP	AI1	AI2	Freq1	MAF	AvgCal l	Rsq	Chr	BP	EFFECT1	OR	STD_ ERR	WALDCHI SQ	PVALUE
rs11819723	A	G	0.92606	0.07394	0.94575	0.36738	10	121717481	-0.329	0.719	0.16	4.2182	0.03999
rs12572375	A	G	0.9306	0.0694	0.94602	0.30797	10	121713430	-0.437	0.646	0.179	5.9374	0.01482
rs12773752	C	T	0.95409	0.04591	0.96746	0.41205	10	121725096	-0.41	0.663	0.189	4.7298	0.02964
rs12776453	C	T	0.96176	0.03824	0.97265	0.4048	10	121690296	-0.434	0.648	0.206	4.4645	0.03461
rs12783135	A	G	0.92798	0.07202	0.94839	0.38542	10	121729817	-0.338	0.713	0.158	4.5477	0.03296
rs12785012	C	T	0.92376	0.07624	0.94403	0.36227	10	121723415	-0.341	0.711	0.159	4.5806	0.03234
rs34407376	G	A	0.95135	0.04865	0.96541	0.40116	10	121714820	-0.408	0.665	0.186	4.7985	0.02848
rs34415434	G	A	0.95555	0.04445	0.96867	0.41896	10	121717173	-0.401	0.669	0.19	4.4598	0.0347
rs77141039	G	A	0.9508	0.0492	0.96502	0.40066	10	121720445	-0.411	0.663	0.185	4.9115	0.02668

**Table 4: Results of 20kb windows between PD initiation and termination of PD GWA hits.**

We viewed windows  $\pm 20$  kilobases around PD GWA study loci derived from Nalls et al 2014 and from the closest genes associated with mRNA expression differences.<sup>169,324</sup> None of the variants within this region demonstrated an association with MSA disease phenotype even upon using a liberal significance cutoff of  $p < 0.05$ . This table includes the most significant variants (labeled by rs number) in these windows.

OR= odds ratio. STD ERR = standard error. WALD = Wald test. CHISQ = chi-squared test.

### 2.3.4 Heritability analysis

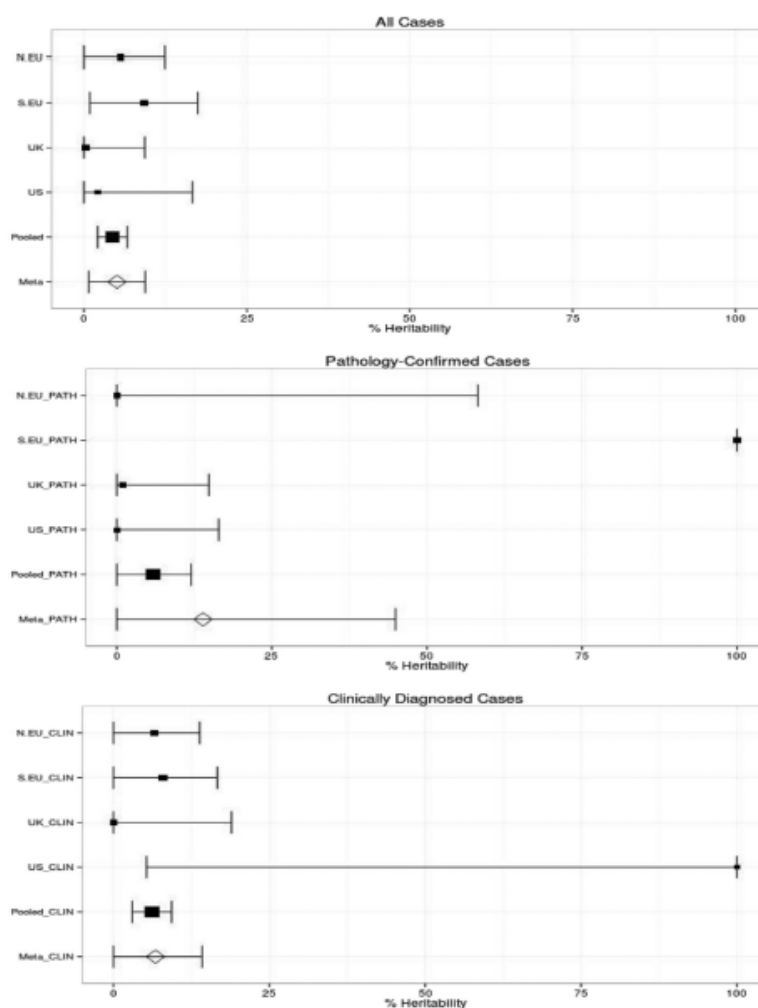
First we used our pooled samples to estimate heritability with GCTA, then divided analyses by population cohort and whether subjects were diagnosed upon autopsy or clinically (Table 5, Table 6, Table 7, Figure 14). Following our heritability estimates of each of these subgroups, we also ran a meta-analysis under a random effects model of all population cohorts in each diagnostic subset: all cases, pathologically-confirmed cases, and cases identified exclusively through clinical diagnosis.

In the pooled MSA sample cohort, we estimated heritability to be about 4.37% in imputed data (95% CI 2.09-6.65%) (Table 5). Looking at specific geographic cohorts, there was a substantial range of estimated heritability, from 0.26% in United Kingdom cases to 9.18% in Southern European cases. While the overwhelming majority of Northern and Southern European samples were identified by clinical means alone, the United Kingdom and United States cohorts were comprised primarily of pathologically-confirmed cases.

Given the high misdiagnosis rate of many parkinsonian disorders, with MSA perhaps being the most renown, pathologically-confirmed cases are significantly more reliable than those cases only receiving clinical diagnoses. Thus, with the intention of minimizing heritability stemming from genetic underpinnings of other neurodegenerative diseases (i.e. PD, PSP, DLB), we performed a separate analysis to estimate the heritability of pathologically-confirmed cases alone (Table 7).

The results of the pooled pathologically-confirmed samples demonstrated an estimated heritability of nearly zero in genotyped data. Intriguingly, however, this estimate rose to around 5.8% (95% CI 0-11.99%) in the imputed data set, suggesting that the imputed genotypes significantly contribute to the heritability of MSA (Table 7). Samples receiving only clinical diagnoses manifested a slightly higher heritability estimate (6.17%) than both the pooled estimate of all cases as well as the pathologically-confirmed cases in the imputed datasets. However, an important caveat to this: both the pathologically-confirmed and clinically diagnosed subgroup heritability estimates are characterized by a fairly large standard error, limiting conclusions that can be drawn (Figure 14, Table 6, Table 7). Moreover, high inter-sample heterogeneity in both the

clinical and pathologically-confirmed subsets (30.2% and 78.6% respectively- $I^2$ ) was revealed in our meta-analyses. Ultimately, this may suggest that our geographic subpopulations have some inter-population genetic differences that cannot be explained by random variation alone.



**Figure 14: Heritability by cohort in diagnostic subgroups**

The size of the center point of these graphs is scaled to the sample size of each subgroup. Some of our cohorts have very high standard errors due to low numbers of cases vs. controls. Pooled = combined results of four geographic subgroups. Meta = Meta-analysis of subgroups under random effects model. Pooled and subgroup cohorts are represented by black squares. Meta-analysis groups are represented by open white diamonds.

(Reproduced from Federoff and Price et al. 2015).<sup>318</sup>

Total: Cases & Controls (N)	Cases: Path (N)	Cases: Clinical (N)	Cohort	Imputation Status: 107,447 Genotyped SNPs / 11,138,628 Imputed SNPs	Heritability	Standard Error	95% Confidence Interval (Percentage Heritability)	P-value
1252	28	258	N.EU All	Genotyped	0.1067	0.0435	2.13-19.21%	0.007724
				Imputed	0.0563	0.0346	0-12.42%	0.04445
1418	14	220	S.EU All	Genotyped	0.1902	0.0506	9.09-28.95%	7.903e-05
				Imputed	0.0918	0.0423	0.88-17.48%	0.01439
920	108	19	US All	Genotyped	0.06652	0.0941	0-25.10%	0.2399
				Imputed	0.0210	0.0742	0-16.65%	0.3858
1183	141	97	UK All	Genotyped	0.0000	0.0545	0-9.33%	0.5
				Imputed	0.0026	0.0463	0-7.63%	0.4793
4784	291	616	All Pooled	Genotyped	0.0490	0.0139	2.1-7.63%	0.0001796
				Imputed	0.0437	0.0116	2.09-6.65%	3.252e-05
			Cohort Meta-Analysis Random Effects Model	Genotyped $I^2 = 55.1\%$	0.0962	0.0424	1.32-17.94%	0.0232
				Imputed $I^2 = 0\%$	0.0506	0.0221	0.72-9.40%	0.0223

**Table 5: Heritability estimate by cohort and subgroup.**

Highlighted text represents the estimated % heritability of imputed genotypes among all pooled cases (4.37%), and the corresponding confidence interval, 2.09-6.65%.  $I^2$  = heterogeneity statistic.

(Reproduced from Federoff and Price et al. 2015).<sup>318</sup>

Total: Cases & Controls (N)	Cases: Path (N)	Cases: Clinical (N)	Cohort	Imputation Status: 107,447 Genotyped SNPs, 11,138,628 imputed SNPs	Heritability	Standard Error	95% Confidence Interval (Percentage Heritability)	P-value
1224	0	258	N.EU Clinical	Genotyped	0.1184	0.0461	2.81-20.89%	0.005081
				Imputed	0.0658	0.0369	0-13.83%	0.03057
1404	0	220	S.EU Clinical	Genotyped	0.1977	0.0536	9.27-30.29%	0.0001884
				Imputed	0.0791	0.0446	0-16.67%	0.03743
812	0	19	US Clinical	Genotyped	1.146861	0.558051	0-100%	0.01377
				Imputed	0.5862	0.4435	5.31-100%	0.089
1042	0	97	UK Clinical	Genotyped	0.0000	0.1177	0-23.09%	0.5
				Imputed	0.0000	0.0965	0-18.93%	0.5
4482	0	616	Clinical Pooled	Genotyped	0.0854	0.0192	4.76-12.32%	4.584e-06
				Imputed	0.0617	0.0161	3.02-9.33%	2.247e-05
4482	0		Clinical Cohort Meta-Analysis Random Effects Model	Genotyped	0.1406	0.0420	5.82-22.31%	0.284
				Imputed	0.0674	0.0382	0-14.23%	0.078
				$I^2 = 30.2\%$				

**Table 6: Clinical cohorts**

The estimated % heritability of imputed genotypes among all clinically confirmed cases is 6.17% with a confidence interval of 3.02-9.33%.

Reproduced from (Federoff and Price et al. 2015).<sup>318</sup>

Total: Cases & Controls (N)	Cases: Path (N)	Cases: Clinical (N)	Cohort	Imputation Status: 107,447 Genotyped SNPs, 11,138,628 imputed SNPs	Heritability	Standard Error	95% Confidence Interval (Percentage Heritability)	P-value
972	28	0	N.EU Pathology Confirmed	Genotyped	0.114253	0.385713	0-87.02%	0.3866
				Imputed	0.000001	0.296941	0-58.2%	0.5
1198	14	0	S.EU Pathology Confirmed	Genotyped	2.269129	0.726325	100%	1.413e-06
				Imputed	2.269129	0.601680	100%	4.586e-07
901	108	0	US Pathology Confirmed	Genotyped	0.0000	0.1076	0-21.09%	0.5
				Imputed	0.0000	0.0839	0-16.45%	0.5
1086	141	0	UK Pathology Confirmed	Genotyped	0.0545	0.0822	0-21.09%	0.2514
				Imputed	0.0100	0.0706	0-14.85%	0.4467
4168	291	0	All Pathology Confirmed	Genotyped	0.0050	0.0375	0-7.85%	0.4467
				Imputed	0.0580	0.0315	0-11.99%	0.03102
4168	291	0	Path-Confirmed Cohort Meta-Analysis: RE Model	Genotyped $I^2 = 68.7\%$	0.1398	0.1615	0-45.62%	0.3865
				Imputed $I^2 = 78.6\%$	0.1391	0.1584	0-44.95%	0.3799

**Table 7: Pathologically confirmed cohorts.**

Highlighted text represents the estimated % heritability of imputed genotypes among all pathologically confirmed cases, 5.80%. The confidence interval, 0-11.99%, is not highlighted due to the limited sample size and very high standard error of the pathologically confirmed cohort. Abbreviations: R.E. = Random Effects model.

(Reproduced from Federoff and Price et al. 2015).<sup>318</sup>

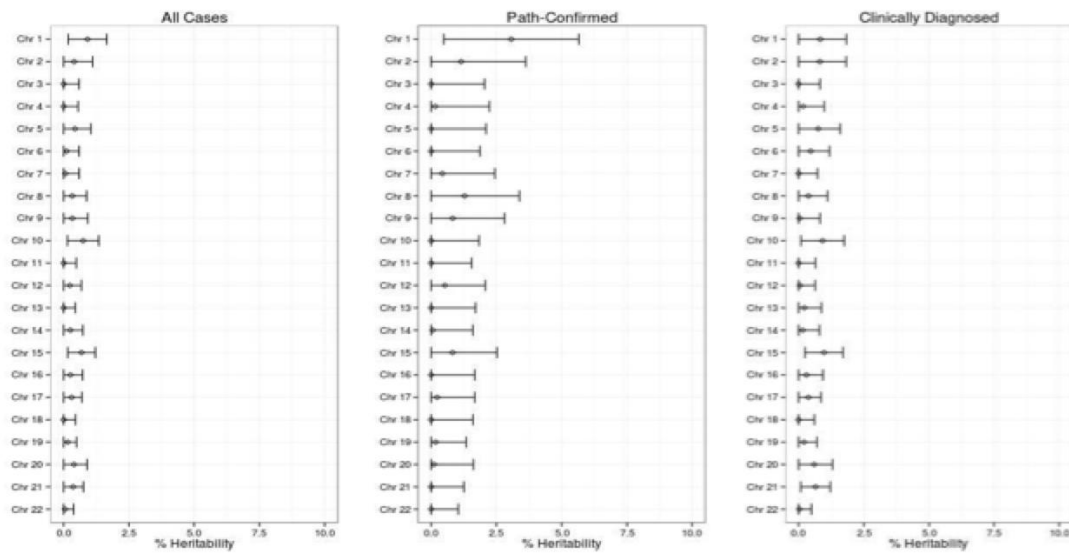


In order to assess which chromosomes contributed most to MSA heritability, we estimated the heritability from each chromosome. In general, individual chromosomal contributions could not be said to contribute more than 0% at 95% confidence, tending to account for less than 1% of total heritability. While chromosome 15 passed multiple-test corrections ( $p < 0.05/22 = 0.00227$ ) in our clinical-only subgroup, with an estimated heritability of 0.25-1.70%, it failed to pass the significance threshold in our pathologically-confirmed only subgroup or our 'all-cases' subgroup. In contrast, chromosome 10 contributed significantly to heritability in our 'all cases' subgroup (0.43-1.05%) but neither in the clinical-only nor pathologically-confirmed only subgroups. Overall, the subgroups including clinical cases revealed higher heritability estimates than those in pathologically-confirmed cases alone for both chromosomes 10 and 15 (Table 8, Figure 15).

Chromosome	Imputation	Heritability	Std.Err	Lower_percent_Range	Upper_percent_Range	P-value	Herit_percent
1	Genotype	0.007203	0.005589	0	1.815744	0.09503	0.7203
1	Imputed	0.008259	0.005123	0	1.830008	0.04984	0.8259
2	Genotype	0.008948	0.005513	0	1.975348	0.04862	0.8948
2	Imputed	0.008131	0.005148	0	1.822108	0.05299	0.8131
3	Genotype	0.006413	0.005194	0	1.659324	0.1095	0.6413
3	Imputed	0	0.004169	0	0.817124	0.4965	0
4	Genotype	0.003621	0.004731	0	1.289376	0.2164	0.3621
4	Imputed	0.001717	0.004137	0	0.982552	0.3404	0.1717
5	Genotype	0.0047	0.004818	0	1.414328	0.1588	0.47
5	Imputed	0.007516	0.004307	0	1.595772	0.03273	0.7516
6	Genotype	0	0.004685	0	0.91826	0.4974	0
6	Imputed	0.004611	0.003697	0	1.185712	0.07451	0.4611
7	Genotype	0.000794	0.004464	0	0.954344	0.4287	0.0794
7	Imputed	0	0.003704	0	0.725984	0.5	0
8	Genotype	0.005226	0.004546	0	1.413616	0.1215	0.5226
8	Imputed	0.003764	0.003735	0	1.10846	0.1301	0.3764
9	Genotype	0	0.004278	0	0.838488	0.4964	0
9	Imputed	0.00045	0.003944	0	0.818024	0.4563	0.045
10	Genotype	0.006845	0.004681	0	1.601976	0.07098	0.6845
10	Imputed	0.009179	0.004218	0.091172	1.744628	0.006315	0.9179
11	Genotype	0.001749	0.004313	0	1.020248	0.3421	0.1749
11	Imputed	0	0.003296	0	0.646016	0.5	0
12	Genotype	0.003083	0.00439	0	1.16874	0.2397	0.3083
12	Imputed	0.000592	0.002947	0	0.636812	0.4127	0.0592
13	Genotype	0.003574	0.003843	0	1.110628	0.1715	0.3574
13	Imputed	0.002191	0.003353	0	0.876288	0.2534	0.2191
14	Genotype	0.002935	0.003656	0	1.010076	0.2073	0.2935
14	Imputed	0.001556	0.003297	0	0.801812	0.3248	0.1556
15	Genotype	0.009746	0.003827	0.224508	1.724692	0.003896	0.9746
15	Imputed	0.009745	0.003708	0.247732	1.701268	0.002077	0.9745
16	Genotype	0.001555	0.003606	0	0.862276	0.3297	0.1555
16	Imputed	0.003013	0.003223	0	0.933008	0.154	0.3013
17	Genotype	0.000632	0.003448	0	0.739008	0.4251	0.0632
17	Imputed	0.003753	0.002468	0	0.859028	0.01234	0.3753
18	Genotype	0	0.003495	0	0.68502	0.5	0
18	Imputed	0	0.003101	0	0.607796	0.5	0
19	Genotype	0.00693	0.003258	0.054432	1.331568	0.01043	0.693
19	Imputed	0.002091	0.002517	0	0.702432	0.1818	0.2091
20	Genotype	0.008739	0.003588	0.170652	1.577148	0.004701	0.8739
20	Imputed	0.005974	0.00358	0	1.29908	0.05058	0.5974
21	Genotype	0.006574	0.002819	0.104876	1.209924	0.006433	0.6574
21	Imputed	0.0065	0.002892	0.083168	1.216832	0.007358	0.65
22	Genotype	0.003942	0.002771	0	0.937316	0.07063	0.3942
22	Imputed	0.000333	0.002331	0	0.490176	0.4449	0.0333

**Table 8: Heritability estimates by chromosome.**

Highlighted values represent imputed chromosomes that were statistically significant only upon the inclusion of clinically diagnosed cases.



**Figure 15: Chromosomal heritability estimates by diagnostic subgroup.**

Heritability estimates of each chromosome are represented by black diamonds. Confidence intervals are illustrated by vertical boundaries.

### 2.3.5 Bayesian estimate of PD-derived heritability

Individuals with diseases such as PSP, PD, and CBD frequently receive a diagnosis of MSA due to the heterogeneous clinical presentation and often irregular disease progression of MSA.<sup>205,322,325,326</sup> As our previous findings illustrate that pathologically confirmed MSA cases have lower estimates of heritability than clinically diagnosed cases, we estimated how much heritability could be expected due to a subset of our clinically diagnosed cases receiving a misdiagnosis of MSA. Our model is based on the following assumptions:

- 1) Based on the 95% confidence interval of the most recent clinical diagnosis positive predictive values from Osaki et al. 2009, MSA false positives comprise approximately 6-25% of our clinical cases.<sup>322</sup> We decided to use this measure

rather than first clinical diagnosis under the assumption that our cases had had several follow-up appointments in order to obtain a thorough clinical history including genotypic data. We assumed that all patients diagnosed post-mortem are true positives.

- 2) Due to the much higher prevalence of PD compared to other atypical parkinsonian diseases (i.e. PSP, CBD) whose clinical phenotype could be mistaken for MSA, we assumed that PD will comprise the overwhelming majority of these false positives. In this instance, we estimated that approximately 70% of misdiagnosed cases would be true PD cases.
- 3) While the heritability of late-onset PD is estimated to be least 31%, we designated a conservatively low heritability estimate of 10% to false positives with diseases other than PD (i.e. CBD, PSP, DLB).<sup>49</sup>
- 4) By assuming the contributions to MSA heritability are additive, we can sum heritability stemming from misdiagnosis of different diseases without considering pleiotropy, which occurs when a single gene influences two or more allegedly disparate phenotypic traits.

Given these assumptions, we calculated the heritability estimate due to misdiagnosis ( $H_m$ ) with the following formula (see methods for more in-depth derivation):

$$H_m = \text{Clin}/\text{Total} * 0.247 * \text{FPR}$$

Where  $H_m$  is the part of the heritability estimate driven by misdiagnosis of other diseases (PD, PSP, CBD, DLB), **Clin** is the number of clinical cases, **FPR** is the false

positive rate, and **Total** is the sum of clinical and path-confirmed cases. Substituting the case statistics for our MSA cohort:

$$H_m = (616 \text{ clinical} / 907 \text{ total}) * 0.247 * \text{FPR} = 0.1677 * \text{FPR}$$

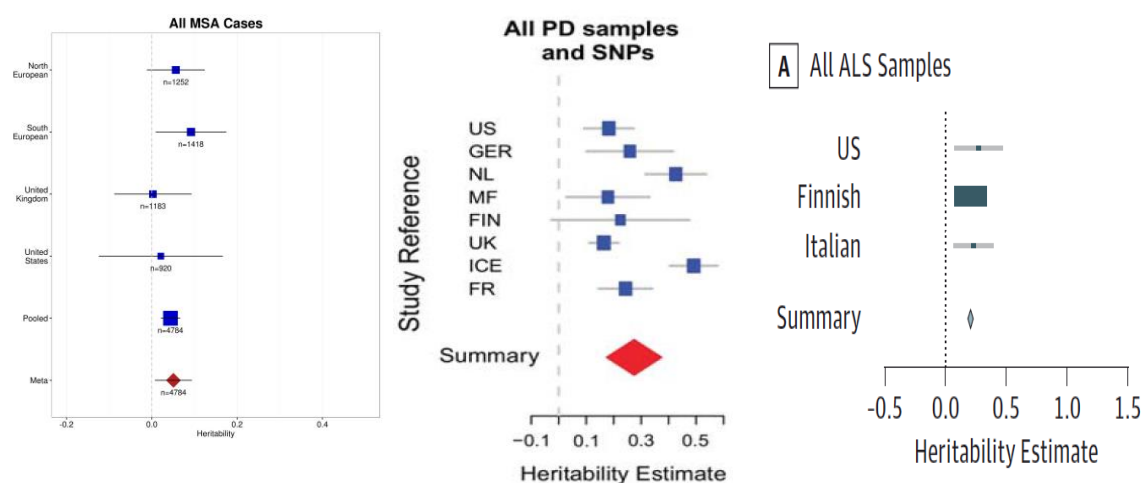
Using the false positive rate of 6-25% derived from Osaki et al 2009<sup>322</sup>, we calculated our expected heritability due to misdiagnosis as **1.00-4.19%**.

## 2.4 Discussion

The CDCV hypothesis, which serves as an impetus to pursue GWA studies, suggests that genetic risk of common diseases are derived, at least partially, from allelic variants with a minor allele frequency (MAF) >1%.<sup>327,328</sup> Though this approach is geared towards elucidating common variants in diseases characterized by a high prevalence such as diabetes mellitus, it can be a useful tactic for studying rare diseases by revealing genes associated with biological and etiological processes. Alternatively, the MRV hypothesis argues that rare variants are liable for the genetic etiology of common, complex diseases.<sup>16</sup> Although former hypotheses suggested a clear dichotomy between CDCV and MRV paradigms, a more profound understanding of genetic architecture now suggests that they occur in tandem, acknowledging the heterogeneous etiology of complex diseases. While we recognize that very rare variants with high penetrance may contribute to the risk of developing MSA, assessing the synergistic effect of common variants associated with MSA will be crucial towards solving the polygenic inheritance puzzle of MSA. By gleaning insight from the latter, we believe this will inform the field both in understanding which genetic approaches are most likely to yield results, and roughly the

amount of genetic influence we can anticipate.<sup>10</sup> As the very first study estimating overall MSA heritability, we aimed to provide a first piece of this puzzle and spark further research to elucidate the genetic risk factors of MSA.

Previous studies using GCTA to estimate heritability of complex neurodegenerative diseases have yielded intriguing results: heritability estimates derived by simultaneously measuring all tagged SNPs have revealed values of 27% and 21% for PD and ALS, respectively (Figure 16).<sup>49,329</sup>



**Figure 16: Disease-specific heritability estimates.**

Heritability estimates are represented by each color, and the shape corresponds to the sample size within that population. Confidence intervals of the summary heritability estimates are demonstrated by horizontal lines associated with each cohort square.

(Reproduced from Federoff and Price et al. 2015, Keller et al. 2012, Keller et al. 2014).<sup>49,318,329</sup>

Furthermore, heritability estimates from GCTA are usually much higher than those estimated from variance in GWA-significant loci alone, as the PD GWA study estimated a heritability of a mere 3%, while ALS was at 12%.<sup>49,329</sup> However, using

GCTA, heritability estimates approach those reported from twin studies.<sup>330,331</sup> The identification of this missing heritability, which can be explained by the inability of GWA study hits to account for the full genetic variance of the underlying phenotype, suggests genetic discoveries yet to be made for diseases like PD and ALS. Moreover, these unidentified genetic variants can be uncovered without possible confounding factors of twin studies, such as shared environment or similar treatment of twins. As GCTA has the ability to combine the small effects of variants not passing significance thresholds in GWA studies, ultimately by analyzing SNPs in a simultaneous fashion, a much more comprehensive yet unbiased assessment of heritability of a particular phenotype is scientifically tractable.

Since MSA demonstrates several overlapping clinical features with both PD and ALS, with an estimated 14% of MSA cases misdiagnosed as other neurodegenerative diseases,<sup>331</sup> it would be reasonable to hypothesize that MSA may reveal a similar heritability estimate using GCTA. Nonetheless, even after using imputed genotypes, MSA heritability estimates are markedly lower than those for PD or ALS: the mean post-imputation MSA heritability was demonstrated to be <10% in all subgroups, with the 95% confidence interval in most subgroups overlapping 0% (Figure 14). Significantly higher estimates of heritability are illustrated in cohorts in which pathologically-confirmed cases comprise a very small proportion of the total population (Northern and Southern European) in comparison to geographic cohorts in which such cases constitute the majority (United States and United Kingdom) (Table 6, Table 7). As the gold-standard for MSA cases due to the known problem of MSA misdiagnosis, pathologically-confirmed samples comprise only a third of our already limited sample size (291

Pathologically-Confirmed / 907 Total). As a result, our attempts to estimate heritability in pathologically-confirmed cases suffer from very high standard errors given this lower sample size and limited statistical power. We acknowledge that our study sample size in conjunction with both the number and distribution of GWA study panel markers limit the lower boundary of variant frequency detected using GCTA. Further, we recognize this would improve by acquiring a greater sample size. It is evident that the limited sample sizes of some cohorts lead to unreasonable heritability estimates after GCTA adjusts for the very low disease prevalence. Ideally, GCTA necessitates large sample sizes of at least several hundred cases to provide reliable estimates. Since part of our analyses utilized very small case cohorts, with some including <100 individuals, heritability calculations derived from these cohorts yielded highly unreliable estimates (i.e. 19 cases in United States clinical cohort, 14 cases in Southern European pathologically-confirmed). This is portrayed by the extensive heterogeneity of our cohort meta-analyses.

While many of our samples were from distinct geographic populations, it was important to consider how much weight to put on heritability from each region. While multidimensional scaling eliminated population outliers in our quality control analyses, there is obviously still some genetic heterogeneity between distinct regional cohorts. Notably, studies in PD have demonstrated GWAs between-quintile odds ratios of a similar magnitude between distinct Caucasian geographic cohorts, suggesting that PD risk profiles of one European location can apply to others within the population stratification boundaries.<sup>332–334</sup> While such studies have only focused on PD, the absence of studies in MSA precludes us from knowing if this same trend applies. However, given the very low prevalence of MSA and high rate of misdiagnosis, the ability to perform



such studies has not been attainable. Thus, as MSA shares pathologic, pathogenic and clinical features with PD, we analyzed our data by applying this same concept, assuming that risk profiles for MSA are likely of similar magnitudes between our 4 geographic cohorts. While we performed individual regional level analyses for both clinically and pathologically confirmed subgroups, our confidence intervals are significantly higher and our statistical power is exceptionally lower than when using our full cohort, making those results much more challenging to interpret.

Chromosomal level heritability estimates demonstrated that the overwhelming majority of chromosomes contribute to almost negligible heritability ( $<1\%$ ) towards MSA, which is not surprising given the overall very low heritability estimates (Table 8, Figure 15). Despite the fact that some chromosomes passed significance cutoffs regarding their genetic contribution to MSA, these findings failed to replicate uniformly across clinically and pathologically diagnosed subdivisions. For example, although chromosome 1 appears to carry a substantial proportion of the heritability in pooled pathologically-confirmed cases, this difference does not pass significance thresholds after multiple testing corrections ( $p < 0.05/22$ ).

Looking at the clinical-only subgroup, chromosome 15 appears to contribute to MSA heritability; however, a trending relationship does not even exist in the pathologically-confirmed subgroup (Table 8, Figure 15). Thus, it is evident that chromosome 15 exhibits a weaker association with the disease after the clinically and pathologically diagnosed cases are pooled together. Given our limited power, it is challenging to say whether this difference is due to biological etiology in pathologically-confirmed and clinical-only subgroups or simply by chance, but this result supports the

notion that some of our estimated MSA heritability may be attributed to clinical misdiagnosis.

Our initial findings may suggest that MSA lacks a substantial common variant heritable component; nonetheless, this is not necessarily indicative of the absence of genetic risk genes and/or variants. There are several factors that can explain the low heritability estimates generated in our study. Primarily, the SNPs incorporated within standard GWA studies are limited to the common variants tagged by microarray-based genotyping methods. If putative causal variants associated with MSA are extremely rare (i.e.  $MAF < 1\%$ ) and consequently not tagged by genotyping platforms, they will be missed. Moreover, if a very rare variant only exists within a single case among the full cohort, it will not be recognized as a shared genotypic variant among cases and thus would not contribute to overall estimation of MSA heritability, as the similarity between cases in the GRM will not increase. Thus, factors associated with more rare variant detection implicitly highlight the essential role of sample size in this type of analysis. Further on this notion, we are lacking the ability to detect rare variants that could explain MSA etiology within an affected family, as our cohort consists of all idiopathic MSA cases which cannot be further scrutinized via segregation and linkage analyses. Though the issue of sample size cannot be altered, the other challenge of variant detection is somewhat ameliorated by imputation; therefore, this case-control study suggests that only modest genotypic variation in common SNPs exists between MSA cases and controls, as imputation incorporates reference haplotypes, suggesting that rarer genetic variants will remain undetected.

Secondly, we must consider the possibility of incomplete LD; hence, even if causal variants are included within haplotype stretches of SNPs in the array, they may be in incomplete LD with the SNPs that have been genotyped.<sup>335</sup> It is important to recognize that this is not mutually exclusive with the former, since causal variants in incomplete LD may likewise exhibit a  $MAF < 1\%$ , further exacerbating these effects.

Thirdly, it is challenging to impute rare variants from array-based genotypes, as genotype platforms are typically defined by common variants and imputation relies on LD. Along with the exclusion of potential novel rare variants, it is also critical to acknowledge that GCTA analysis does not account for non-additive genetic factors (i.e. epistasis) and possible environmental effects. In essence, this implies that the heritability estimate calculated by GCTA defines a lower limit of MSA heritability that would likely increase if such factors could be integrated accordingly.

Lastly, the distinctions in diagnostic status (clinically vs. pathologically-confirmed) require further scrutiny. While the Southern and Northern European cohorts consisted of the highest number of clinically diagnosed MSA cases and very few pathological cases, the United States and United Kingdom together comprised 82% of all pathologically-confirmed cases. Intriguingly, these geographic subset differences resonate with estimated heritability levels: the United States and United Kingdom cohorts demonstrate lower heritability estimates, ranging between 0 and 3%, while the Northern and Southern European cohort estimates are much significantly higher, ranging between 0 and 17.48% (Figure 14, Table 5) Given the 6-25% clinical misdiagnosis rate of MSA, this is particularly noteworthy, suggesting that a substantial proportion of those cases diagnosed as MSA are indeed PD cases, as late-onset PD exhibits a higher heritability

estimate of about 31%.<sup>49,322</sup> Taken into consideration, it is quite plausible that misdiagnosed cases within our clinical subpopulation (in all geographic cohorts, but particularly those consisting of predominately clinical cases) are inflating the heritability estimate of MSA.

Upon reflecting back on our calculated Bayesian estimate of heritability due to misdiagnosed MSA, ranging between 1.00-4.19%, there is substantial overlap with our GCTA estimate, ranging between 2.09-6.65% (Table 5). As we derived our calculations based upon a comprehensive literature overview of misdiagnosis rates, the overlap of these estimates suggest that all MSA heritability estimated in this study could in principle be explained exclusively through heritability stemming from contamination of the MSA cohort with non-MSA diseases.

Such a result highlights the multitude of challenges in attempting to discover genetic risk factors for MSA: first, as the prevalence of MSA is incredibly low, estimated at approximately 0.000046, sample size is rather limited, and many cases that are clinically diagnosed are likely misdiagnosed, adding noise to any genetic variation that may underlie MSA etiology.<sup>320,321</sup> While an obvious approach to improve the relevance and validity of MSA genetic analyses would be to include only pathologically-confirmed cases, any such attempts would necessarily be underpowered due to the rarity of the disease. Despite the fact that our dataset represents the most comprehensive collection of MSA genotypes ever assembled, our cohort numbers under 1000 cases lacks adequate statistical power necessary to detect uncommon variants and/or those with mild effects. It is thus in the scientific community's best interest that we seek international collaboration to generate large, high-confidence (i.e. pathologically-confirmed), high-quality datasets.

Moreover, given our findings and acknowledged limitations of this methodology, the use of NGS technology in pursuit of MSA genetic etiology could prove extremely valuable, as exon-centric variation may reveal novel rare variants that have been missed by standard genotyping methods used in this investigation.

### **3 Identifying candidate genes and variants for MSA using exome sequencing**

*Statement of Contributions:* Collection of the MSA samples was performed by Dr. Lucia Schottlaender and Dr. Henry Houlden. Dr. Schottlaender and I worked together on the exome sequencing of the MSA samples. She spent time at the Laboratory of Neurogenetics, NIH to prepare and run 200 pathologically confirmed samples using the Illumina Tru Seq protocol. I prepared and ran 212 clinically confirmed MSA samples using the Illumina Nextera protocol. The details of both protocols will be discussed in the methods section of this chapter. I performed quality control and data analysis, under the supervision of the Statistical Genetics Group and the Computational Biology Core of the Laboratory of Neurogenetics.

#### **3.1 Introduction**

In our pursuit of unraveling the genetic etiology of MSA, we recognize the possibility of both common and rare variants affecting the risk for disease based on the CDCV and MRV hypotheses. In the first chapter, we investigated the role of common variants associated with disease to estimate MSA heritability defined by common variation alone. Had our MSA heritability estimates mirrored those of other neurodegenerative diseases like PD and ALS, we would have investigated particular loci to identify such variants or genes. However, as the MSA heritability estimate based on common variation is between 4-5%, we believe common variation does not play a

substantial role in risk or association with disease. While we acknowledge several limitations inherent in genotyping technology and GCTA analysis, suggesting that we interpret our results with caution, the data indicates that the MRV hypothesis may be more applicable to a very rare disease like MSA.

In an effort to identify rare variants associated with disease, a logical nexus is to pursue WES. Within the last several of years, second generation sequencing, and in particular WES has revolutionized the world of genetics. The exome consists of roughly 180,000 exons within approximately 27,000 genes and represents all protein-coding variants in the genome. While the exomic region physically comprises a mere 1-2% of the genome, approximately 85% of human monogenic diseases are caused or associated with missense mutations.<sup>9</sup> As WES yields coverage in the majority of exons within the coding region of the genome, we can identify novel nucleotide variants in the form of missense, nonsense, frameshift, and indel mutations and assess their association with disease through individual variant and gene burden analyses. In the context of MSA, a substantial proportion of these samples should be pathologically confirmed, given the estimated 14% clinical misdiagnosis rate, which could significantly confound results.<sup>322</sup>

Because MSA is a rare disease and because clinical diagnosis is imprecise it is difficult to ascertain a sample size of well-characterized MSA patients of a similar magnitude to that used in other neurodegenerative diseases such as PD or AD. Clearly then, any MSA cohort in current existence is unlikely to be of sufficient depth to provide compelling and replicated genome wide associated variants. However, as has been seen with PD, production of early hypothesis generating datasets spurs investigation and, with public release of results, catalyzes independent replication. With this in mind we chose to

pursue WES in a cohort of clinically and pathologically diagnosed MSA samples, with the express intent of generating a list of candidate associations for this disorder.

## 3.2 Materials and methods

### 3.2.1 Subjects

The MSA cohort consists of apparently sporadic cases with no family history information on relatives with PD or other neurodegenerative disorders. A total of 411 MSA samples were obtained for WES with the majority of individuals from the United Kingdom, France or the United States. The remaining samples were all of other European descent, including samples from Germany, Spain, and the Netherlands. Among these, 212 samples were pathologically confirmed and the remaining 199 received clinical diagnoses of MSA (**Table 9**). The percentage of MSA cohort samples from each country is illustrated in **Figure 17**.

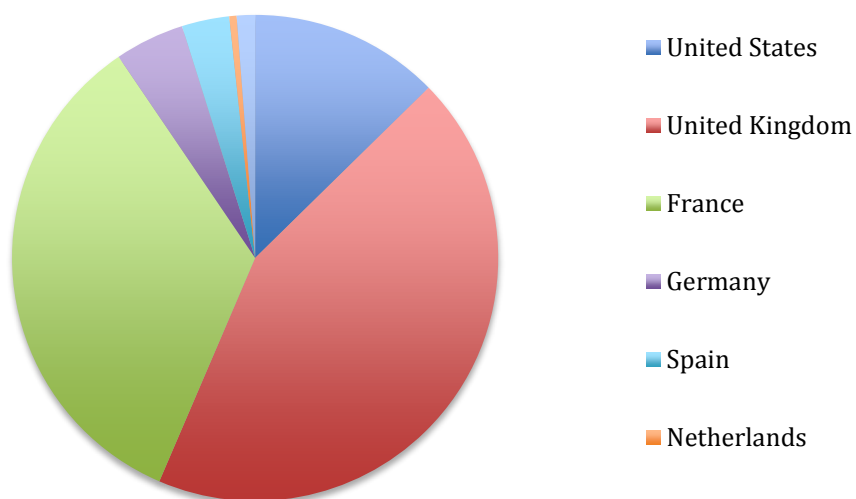
<b>MSA Sample country of origin</b>	<b>Number of samples from country</b>	<b>% Pathologically confirmed</b>	<b>% Clinically diagnosed only</b>
United States	52	100%	-
United Kingdom	180	124/180=69%	56/180 = 31%
France	140	-	100%
Germany	19	100%	-
Spain	13	100%	-
Netherlands	2	100%	-
Unknown European country	5	100%	-

**Table 9: Descriptive statistics of MSA WES cohort**

Information about the country of origin and clinical or pathological diagnostic status was obtained for almost every sample.



## Percentage of MSA cohort samples by country



**Figure 17: Origin of MSA cohort samples**

Visual graphic of table 9.

Gender information was available for 337 individuals, consisting of 177 males and 160 females, and was unavailable for the remaining 74 patients. Records of age of onset of disease were available for 396 patients, while only 139 samples had disease duration information reported. Samples were collected from several locations within each geographic region, as illustrated in Table 10. While the disparate provenance of these samples is not necessarily ideal for a genetic analysis, the rarity of this disorder requires international collaboration in order to gather a sufficient number of samples.

Country of origin	University and/or Hospital	Number of Samples from location
United States	Center for Neurodegenerative Disease Research, University of Pennsylvania	25
United States	University of Miami Brain Bank	10
United States	Emory University Brain Bank	2
United States	Harvard University Brain Bank	2
United States	Johns Hopkins University-Juan Troncosco laboratory	13
United Kingdom	Queen Square Brain Bank (QSBB), University College London	10
United Kingdom	The Manchester Brain Bank, University of Manchester	2
United Kingdom	Newcastle Brain Tissue Resource, Newcastle University	6
United Kingdom	Institute of Psychiatry Brain Bank, King's College London	5
United Kingdom	UK Parkinson's disease tissue bank at Imperial College London	3
United Kingdom	Other (unknown)	154
France	Unknown	140
Germany	Neurobiobank München, Institut für Neuropathologie, Ludwig-Maximilians-Universität, Munich	17
Germany	Brain Bank Center Würzburg	2
Spain	Neurological Tissue Bank, University of Barcelona, Hospital Clinic, Barcelona	10
Spain	Other/unknown	3
Netherlands	Netherlands Brain Bank, Netherlands Institute for Neuroscience, Amsterdam,	2

**Table 10: Origin of samples by contributing center**

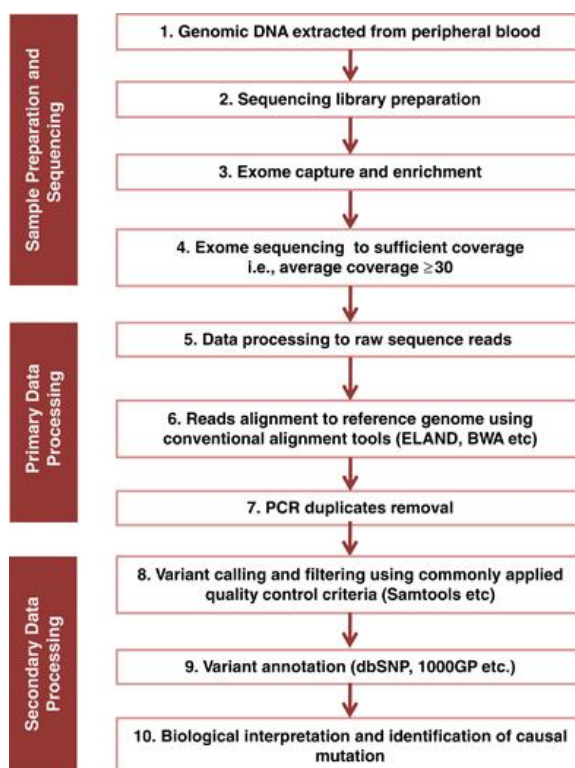
The number of samples obtained from each University and/or Hospital is listed by respective country.

All neurologically normal control samples were obtained from two centers: the Alzheimer's Disease Genetic Consortium (ADGC) and the Cohorts for Heart and Aging Research in Genomic Epidemiology (CHARGE) consortia. The ADGC consortium currently contains ~10,000 control samples while CHARGE consists of more than 43,000

control samples (<https://www.niagads.org/adsp/content/study-design>). All controls are of European American ancestry with respective gender and age information available.

### 3.2.2 Whole exome sequencing

The process of WES consists of four distinct phases, the first three in the form of bench work in the laboratory (Sample Preparation and Sequencing), and the last step (Primary Data Processing, Secondary Data Processing) requiring computationally intensive work using the command line interface, most often on a Linux based computer system (Figure 18).



**Figure 18: Overview of WES pipeline.**

WES can be subdivided into several stages including: sample preparation and sequencing, primary data processing, and secondary data processing.

(Reproduced from Ku et al 2013).<sup>336</sup>

A brief overview of each step in the exome sequencing process and analysis, followed by a more in-depth description of the protocol, is detailed below. These apply to both the Illumina Truseq and Nextera protocols. It is important to recognize that different exome capture kits may vary in the capture efficiency and the specific regions that are covered. Notably, while both Truseq and Nextera protocols both cover 45Mb of exonic content and have at least 80% on-target-sequencing reads, Nextera is approximately 70% faster for library preparation. While the primary goal is to maximize capture of coding sequences, there is typically some capture of introns, untranslated regions (UTRs), and regions encoding non-coding RNA in both protocols.

#### ***3.2.2.1 DNA library prep & enrichment***

Samples are prepared 96 at a time using the Illumina enrichment kit. This consists of several master-mixed reagents, optimized index adaptors, and quantification methods through fluorescent dyes (as opposed to using an agarose gel). Samples are labeled by two distinct indices and pooled into batches of 12 at a time. Each pool can then undergo a clustering preparation protocol to prepare for the next step, cluster generation.

#### ***3.2.2.2 Cluster generation***

The clustering is performed by an automated device called a Cluster Station (c-Bot) and takes place on the surface of a flow cell (FC). The FC is an 8-channel sealed glass micro fabricated device that uses DNA polymerase for the ‘bridge amplification’ of the DNA fragments on its surface, producing multiple DNA copies or clusters. Individual libraries may be run singly or in combination with others (pooled libraries). Each cluster contains approximately one million copies of the original fragment that is sufficient for accurate signal intensity detection during sequencing (Figure 23).

### 3.2.2.3 *Parallel sequencing by synthesis*

All four nucleotides with DNA polymerase are added simultaneously to the FC channels. This is the premise for a sequencing by synthesis approach (Figure 24). The nucleotides carry a base-unique fluorescent label and the 3'-OH group is chemically blocked. Thus, each base incorporation is a unique event that is captured by an imaging step. The 3' blocking group is then chemically removed, preparing each strand for the next base incorporation. This series of steps continues for a specific number of cycles, as determined by user-defined instrument settings, which permits discrete read lengths of 50–100 bases. To create paired-end reads, both strands of DNA undergo identical sequencing by synthesis processes as described above, which plays a key role in both the precision and accuracy of mapping as well as the identification of small structural variants (i.e. indels).

### 3.2.2.4 *Data analysis*

Data processing can be divided in 3 main steps:

First, raw read data are transformed into a single, generic representation, mapped to their genomic origin and aligned consistently. Next, molecular duplicates are eliminated and initial alignments are refined (Figure 18, Primary Data Processing).

Secondly, the analysis-ready SAM/BAM files permit discovery of all sites with statistical evidence for an alternate allele present among the samples (including SNPs and small indels). Next, raw variant calls are integrated with technical covariates, known sites of variation, genotypes for individuals, linkage disequilibrium (LD), and family and population structure. This process enables quality-based scoring of variants as polymorphic sites or artifacts (Figure 18, “Secondary Data Processing”).

Finally, high-quality genotypes are determined for all samples and, after initial mapping and duplicate checking, all samples are run through the Genome Analysis Toolkit (<https://www.broadinstitute.org/gatk/>).

### **3.2.3 Illumina TruSeq protocol**

#### ***3.2.3.1 DNA Library preparation and enrichment***

##### **3.2.3.1.1 Quantification**

Genomic DNA (gDNA) is quantified using the Qubit fluorimetric quantitation system, dsDNA BR Assay kit. 1 $\mu$ g is required. The total volume of DNA needed for the next step (fragmentation) per sample is 52.5  $\mu$ l. Based on the results of Qubit quantification, some samples are vacuumed to reduce volume (using the SpeedVac concentrator) while others require the addition of re-suspension buffer (RSB) to bring the total volume up to 52.5  $\mu$ l. Once each sample reaches this volume with a minimum of 1 $\mu$ g, it is ready for fragmentation.

##### **3.2.3.1.2 Fragmentation of gDNA**

Each sample is placed in Covaris tubes and randomly sheared using the Covaris E210 water bath sonicator. The conditions of the machine are as follows: Duty Cycle: 10%, Intensity: 5, Cycles per Burst: 200, Time: 120 s, Mode: Frequency sweeping, Power 23W, Temperature 5.5°C to 6°C).

##### **3.2.3.1.3 Quality check using the Bioanalyzer**

To assess the quality of the shearing, 1 $\mu$ l of each fragmented dsDNA sample is run on the bioanalyzer using an Agilent DNA 1000 chip. This is important to determine both the size and concentration of sheared DNA fragments. Fragment size is specific to

the type of next generation sequencing method that is utilized. In WES, fragment sizes are approximately 250bp in length, which is necessary to cover the majority of targeted exons (typically 200bp in length).

The protocol for using the Agilent DNA1000 Bioanalyzer is as follows:

***A. Prepare Gel Dye Mix:***

- i. Allow the DNA dye concentrate (blue) and DNA gel matrix (red) to equilibrate to room temperature for 30 minutes.
- ii. Vortex the blue- capped DNA dye concentrate (blue) for 10 seconds and spin down. Make sure the DMSO is completely thawed.
- iii. Pipette 25 ul of the blue capped dye concentrate (blue) into a red- capped DNA gel matrix vial (red). Store the dye concentrate at 4 °C in the dark again.
- iv. Cap the tube, vortex for 10 seconds. Visually inspect proper mixing of gel and dye.
- v. Transfer the gel-dye mix to the top receptacle of a spin filter.
- vi. Place the spin filter in a microcentrifuge and spin for 15 minutes at room temperature at  $2240\text{ g} \pm 20\%$  (for Eppendorf microcentrifuge, this corresponds to 6000 rpm).
- vii. Discard the filter according to good laboratory practices. Label the tube and include the date of preparation.

***B. Loading the Gel-Dye Mix***

- i. Allow the gel-dye mix to equilibrate to room temperature for 30 minutes before use.

Protect the gel-dye mix from light during this time.

- ii. Take a new DNA chip out of its sealed bag and place the chip on the chip priming station.
- iii. Pipette 9.0 ul of the gel- dye mix at the bottom of the well marked.
- iv. Set the timer to 60 seconds, make sure that the plunger is positioned at 1 ml and then close the chip priming station. The lock of the latch will click when the Priming Station is closed correctly.
- v. Press the plunger of the syringe down until it is held by the clip.
- vi. Wait for exactly 60 seconds and then release the plunger with the clip release mechanism.
- vii. Visually inspect that the plunger moves back at least to the 0.3 ml mark.
- viii. Wait for 5 seconds, then slowly pull back the plunger to the 1 ml position.
- ix. Open the chip priming station.
- x. Pipette 9.0 ul of the gel- dye mix in each of the wells marked.

### ***C. Loading the Marker***

- i. Pipette 5 ul of green- capped DNA marker (green) into the well marked with the ladder symbol and into each of the 12 sample wells.

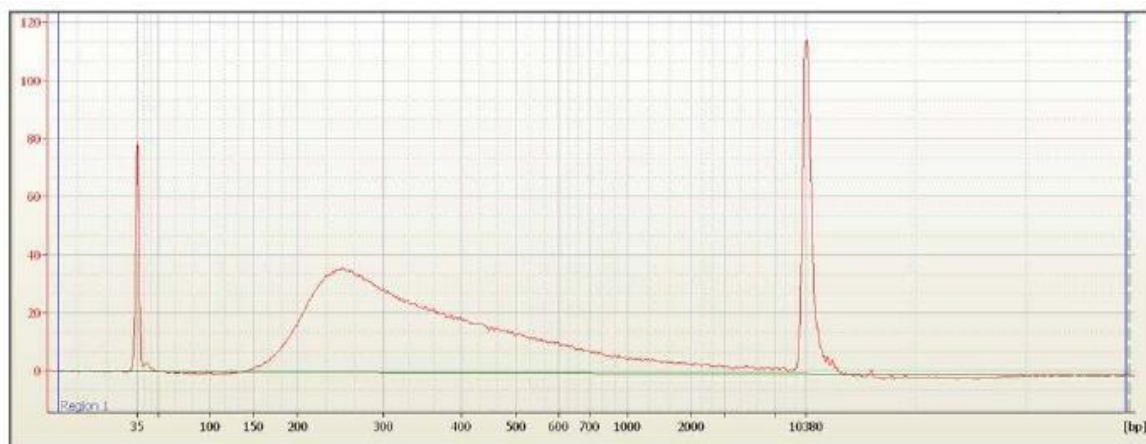
### ***D. Loading the Ladder and Samples***



- i. Pipette 1 ul of the yellow-capped DNA ladder (yellow) in the well marked with the ladder symbol.
  - ii. In each of the 12 sample wells pipette 1 ul of sample (used wells) or 1 ul of deionized water (unused wells).
  - iii. Set the timer to 60 seconds.
  - iv. Place the chip horizontally in the adapter of the IKA vortex mixer and make sure not to damage the buldge that fixes the chip during vortexing.
  - v. Vortex for 60 seconds at 2400 rpm.
  - vi. Refer to the next topic on how to insert the chip in the Agilent 2100 bioanalyzer.
- Make sure that the run is started within 5 minutes.

\_\_\_End Agilent DNA 1000 Chip Protocol \_\_\_

Once the chip has been run, one must review the ladder electropherogram to determine that there are 13 well-resolved peaks with a flat baseline and correct identification of both markers. The electropherograms of each sample must be analyzed individually to determine concentration and fragment size. An example of a successful sample run is shown below.



**Figure 19: High quality library sample on the Agilent bioanalyzer.**

Peaks on each end denote markers. The sample is illustrated by the middle peak, with the majority of fragments around 250bp in length. The x-axis reflects the number of basepairs and the y-axis denotes the fluorescence units as detected by the bioanalyzer.

#### 3.2.3.1.4 Post fragmentation end repair

Following fragmentation and analysis with the bioanalyzer, each fragment has a 3' overhang. These overhangs are transformed into blunt ends using the End Repair Mix (ERM) in the Illumina Truseq Kit. This is accomplished by adding a 3' to 5' exonuclease, which eliminates the 3' overhang, followed by a DNA polymerase, which fills in the remaining 5' overhang. To prepare each 50 ul sample of fragmented DNA, 10 ul RSB and 40ul of ERM are added and the final 100 ul solution is incubated in a thermal cycler for 30 minutes at 30°C.

#### 3.2.3.1.5 Cleaning with AMPure Beads XP

Following end repair, each sample must be cleaned with paramagnetic Ampure beads. This first requires dilution of the beads by combining 125 ul of beads with 35 ul of deionized molecular grade free water. The diluted bead mix (160 ul) is then combined

with each 100 ul sample, followed by mixing and incubation at room temperature for 15 minutes. Following incubation, the 96 deep well plate is placed on a magnetic stand for 15 minutes to adequately let the paramagnetic beads bind the DNA fragments. Upon sufficient binding, 255 ul of supernatant is discarded. Next, while keeping the plate on the magnetic stand, two sequential washes are conducted using a solution of 80% ethanol. After the second wash, the ethanol is removed and discarded and the samples are incubated at room temperature for 15 minutes to sufficiently dry. To re-suspend the DNA, each sample is eluted with 17.5 ul of RSB.

#### 3.2.3.1.6 3' End Adenylation

After cleaning and resuspension, a single 'A' nucleotide is added to the 3' ends of blunt fragments in order to prevent ligation with complementary strands before the upcoming adaptor ligation reaction. The adaptor has a complementary 'T' nucleotide on the 3' end which serves as the corresponding overhang for ligation with the fragment. Each sample is combined with 12.5 ul A-Tailing mix and 2.5 ul RSB, followed by mixing and incubation at 37°C on a thermal cycler for 30 minutes.

#### 3.2.3.1.7 Adapter Ligation

Each sample is combined with 2.5 ul DNA Adapter Index, 2.5 ul Ligation mix and 2.5 ul RSB in order to add indexing adapters to the ends of DNA fragments. This solution is mixed and incubated at 30°C on the thermal cycler for 10 minutes. After incubation, ligation is terminated upon the addition of 5 ul Stop Ligation Buffer to each ligated sample. Samples are then cleaned using Ampure Beads XP, in the same process as described in 3.2.3.1.5.

Subsequently, 1 ul of each sample, now a ligated library, is run on the bioanalyzer using an Agilent DNA 1000 chip, as discussed in detail in 3.2.3.1.3 above. This is necessary to check if the adaptor ligation is successful, as one should visualize DNA fragment lengths (x-axis) corresponding to 400-500 bp in size. The y-axis reflects fluorescence units (FU), which is proportional to the DNA concentration of each sample.

#### 3.2.3.1.8 DNA library enrichment

The goal of this step is to amplify DNA fragments ligated with adapter molecules on each end using PCR. Each sample is mixed with 25 ul PCR Master Mix and 5 ul PCR Primer Cocktail and incubated on the thermal cycler according to the following conditions: 30 seconds at 98°C, 10 cycles of: 10 seconds at 98°C, 30 seconds at 60°C, 30 seconds at 72°C, 5 minutes at 72°C, then hold at 4°C. The PCR products are purified with Ampure XP Beads (refer to section 3.2.3.1.5 for details of cleaning process).

Purification is followed by library quality assessment using both the Agilent DNA 1000 chip on the Bioanalyzer and the Qubit fluorimetric quantitation system (refer to step I for details). Successful bioanalyzer results should reveal a 5-fold increase in peak height (measured by FU), indicative of an increase in DNA concentration and successful amplification.

#### 3.2.3.1.9 Exome Capture

During this step, Illumina TruSeq capture probes are used to capture the adapter-enriched DNA sample libraries prepared in sections 3.2.3.1.1-3.2.3.1.8 above. Using quantification values from section 3.2.3.1.7 above. 500ng of each DNA library is mixed with 500 ng of 11 other unique DNA libraries to make a single 40 ul pool consisting of

12 DNA libraries and containing 6ug of total DNA. Many samples often require further concentration to achieve the correct volume and amount of DNA. Thus, they can be vacuumed on the SpeedVac concentrator to decrease sample volume without the addition of heat. As each library has its own unique indices, the pooling of samples will not hinder the parsing out (“de-multiplexing”) of individual sample sequences in the analysis process.

#### 3.2.3.1.10 First Hybridization

Following exome capture and pooling, samples are prepared for the first hybridization by mixing each 40 ul library pool (consisting of 12 samples) with 10 ul Capture Target Oligos and 50 ul Capture Target Buffer 1. This mixture is incubated on a thermal cycler according to the following conditions: 10 minutes at 95°C, 1 minutes at 93°C for 18 cycles, decreasing 2°C per cycle, followed by 16-20 hours at 58°C.

#### 3.2.3.1.11 First Wash

Immediately following the first hybridization, samples undergo the first washing process to capture probes bound to target exons using Streptavidin Magnetic beads. In a series of three subsequent washes, DNA fragments that are not bound to the magnetic beads are discarded. Specifically, this is accomplished by the addition of 250 ul of Streptavidin Magnetic beads to each pool of hybridized DNA libraries. This solution is thoroughly mixed and then incubated for 30 minutes at room temperature.

After incubation, the deep well plate is put on the magnetic stand for 2 minutes, allowing the unbound DNA fragments to remain in the supernatant, which is subsequently removed and discarded. Next, the deep well plate is removed from the

magnetic stand and 200 ul of Wash Solution 1 is added to each pool and mixed thoroughly. The plate is then put back on the magnetic stand for incubation at room temperature for 2 minutes. Once again, the supernatant is removed and discarded and the plate is taken off the magnetic stand. In a similar fashion, Wash solution 2 is added to each pool and thoroughly mixed to completely re-suspend the magnetic beads. The plate is once again placed on the magnetic stands for 2 minutes to allow the magnetic beads bound to the captured DNA to separate from unbound DNA fragments. The supernatant is then removed and the plate is taken off of the magnetic stand. Finally, 200ul of Wash Solution 3 is added to each pool and mixed thoroughly. This is followed by incubation of the plate on the thermal cycler at 42°C for 30 minutes. Immediately after this step, the plate is returned to magnetic stand for 2 minutes, followed by supernatant removal. To ensure thorough cleaning, the step using Wash Solution 3 is repeated for a second time.

Upon completion of the second round of washing using Wash solution 3, the plate is placed on the magnetic stand and the supernatant is removed and discarded. Next, a 30 ul elution pre-mix, consisting of 1.5 ul NaOH and 28.5 ul Elute Target buffer, is added to each pool and mixed thoroughly to ensure bead re-suspension. This is followed by incubation at room temperature for 5 minutes and then placement on the magnetic stand for 2 minutes. 29 ul of supernatant from each pool is transferred to a new plate and combined with 5 ul of Elute Target Buffer II to form clean, hybridized pool(s) of libraries.

#### 3.2.3.1.12 Second Hybridization

The Second Hybridization is identical to the First Hybridization in (see section 3.2.3.1.10 for details). The goal is to further enhance the enrichment of targeted exonic

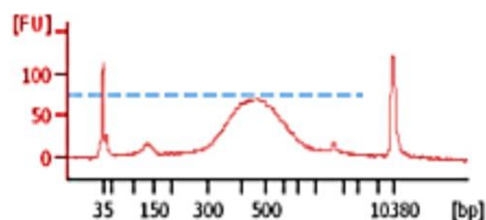
regions by mixing the first round of eluted DNA libraries with the capture probes.

### 3.2.3.1.13 Second Wash

The second wash is identical to the first wash.

### 3.2.3.1.14 Library Enrichment

Upon completion of the second wash, the hybridized library is enriched using the same protocol as section 3.2.3.1.8. The only difference is that the final holding temperature on the thermal cycler is 10°C (instead of 4°C). The amplified pools (each containing 12 libraries) are then washed with Ampure Beads XP using the same procedure as step 3.2.3.1.5. Finally, 1 ul of each pooled sample is bioanalyzed using an Agilent DNA High Sensitivity Chip (as opposed to DNA 1000 chip) to maximize concentration and quality accuracy of each pooled set of samples. To ensure a successful second hybridization and final enrichment, DNA peaks should range from 50-80 FU (Figure 20).



**Figure 20: Electropherogram of a successful library.**

Example of successful bioanalyzer electropherogram using an Agilent DNA High Sensitivity chip following the final step of Library enrichment in the Illumina Truseq protocol. The x-axis reflects the number of basepairs and the y-axis denotes the fluorescence units (FU) as detected by the bioanalyzer. Ladders are represented by peaks around 35bp and 10380 bp and the DNA sample is illustrated as the middle peak with the largest amount of sample being approximately 475 bp in length and 75 FU.

### **3.2.4 Illumina Nextera rapid capture protocol**

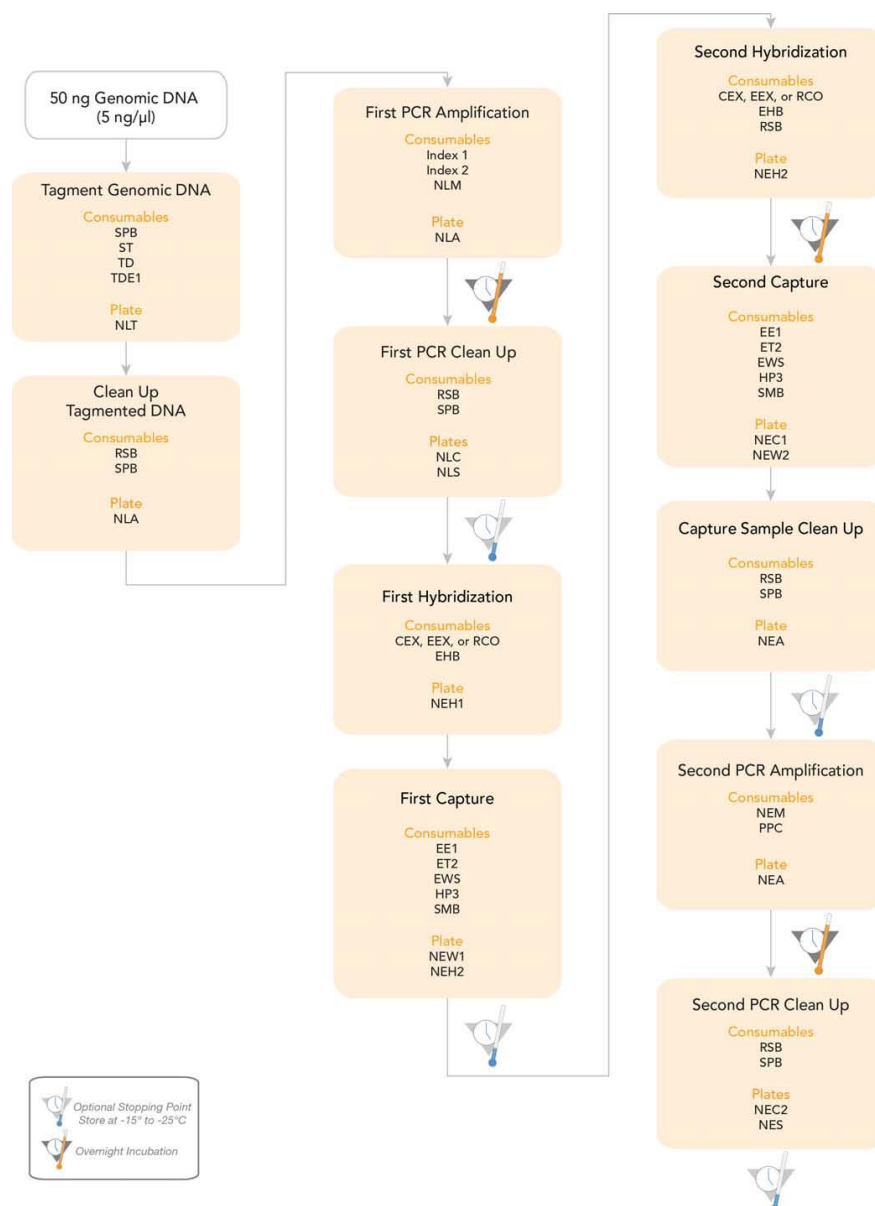
Many of the steps are similar if not identical to the TruSeq protocol. Unique steps will be discussed in detail.

#### ***3.2.4.1 DNA Library preparation and enrichment***

##### **3.2.4.1.1 Quantification**

The first step involves DNA library preparation using the Illumina Nextera Rapid Capture Enrichment Kit. This requires a minimum of 50ng of genomic DNA, at concentration of 5ng/ul. Notably this is a significantly lower amount of DNA required than the Illumina TruSeq protocol previously discussed. DNA is quantified with the Qubit fluorimetric quantitation system as described in section 3.2.3.1.1. A flow chart of the full Nextera library preparation is depicted below in Figure 21.





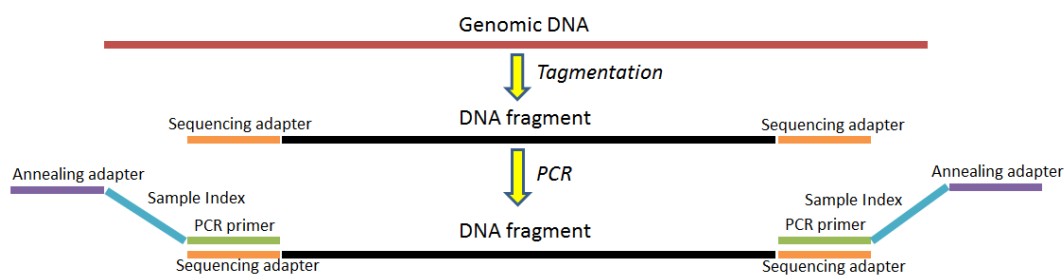
**Figure 21: Nextera rapid capture enrichment process.**

A 50ng sample of Genomic DNA is required to perform WES using the Nextera protocol. Core steps of sample processing include: tagmentation, hybridization, amplification, capture and clean up.

(Reproduced from [www.illumina.com](http://www.illumina.com))

### 3.2.4.1.2 Tagmentation of gDNA

Each sample must undergo a tagmentation process to oligonucleotide adapters, followed by cleaning with Sample Purification Beads (SPB). The process of tagmentation is shown below in Figure 22. This requires a mixture of 10 ul gDNA, 25 ul Tagment DNA Buffer and 15 ul Tagment DNA Enzyme I in a 96-well MIDI plate. *The plate is then placed on a microplate shaker at 1800 rpm for 1 minute, followed by centrifugation at 280 xg for 1 minute (NOTE: these (in italics) are considered standard conditions for this protocol and will be referred to as such from now on. If conditions are different they will be specified).* Next, the plate is incubated at 58°C for 10 minutes. 15 ul of Stop Tagment Buffer is then added to each sample followed by shaking and centrifuging at the standard conditions. Finally, the plate is incubated at room temperature for 4 minutes.



**Figure 22: Tagmentation followed by first PCR.**

Tagmentation allows for sequencing adapters to be placed on both ends of the genomic DNA, followed by subsequent binding of the unique dual indices on each end, thus making each DNA library distinct. This is followed by first PCR using Illumina Nextera Rapid Capture Enrichment Kit.

(Reproduced and modified from Head et al 2014 and Kara et al 2014).<sup>333</sup>

### 3.2.4.1.3 Clean up Tagmented DNA

This cleaning process will be described in detail here and will be referenced in later sections, as it is repeated throughout the Nextera protocol. First, 65 ul of magnetic cleaning beads (SPB) are added to each sample in a deep well MIDI plate followed by

shaking and centrifuging at standard conditions. Next, the plate is put on the magnetic stand for 2 minutes (or until the liquid appears clear). All supernatant is then removed and discarded and 200  $\mu$ l of 80% ethanol solution is added to each well without disturbing the beads. After waiting 30 seconds, the 80% ethanol is removed and discarded. This step is repeated a second time for thorough cleaning and any remaining ethanol must be removed without disturbing the beads. The plate is left to dry on the magnetic stand for 10 minutes and then removed. 22.5  $\mu$ l of RSB are added to each sample well, followed by shaking at standard conditions for 1 minute. The plate is then incubated at room temperature for 2 minutes and then centrifuged for 1 minute at standard conditions. The plate is then put back on the magnetic stand for 2 minutes (or until liquid appears clear). Finally, 20  $\mu$ l of clear supernatant is transferred from each well to a new standard 96 well plate.

#### 3.2.4.1.4 First PCR Amplification

The first PCR amplification process is performed after each sample is tagged with two distinct series of indices (Figure 22, “Sample Index”). The PCR mixture requires 5  $\mu$ l Index I primer, 5  $\mu$ l Index II primer, and 20  $\mu$ l Nextera Library Amplification Mix (NLM) added to each well containing 20  $\mu$ l of sample. The solution will then undergo shaking and centrifugation at standard conditions. Next, the NLM\_AMP program is run on the thermal cycler according to the following conditions: Choose the pre-heat lid option and set to 100°C, 72°C for 3 minutes, 98°C for 30 seconds, 10 cycles of: 98°C for 10 seconds, 60°C for 30 seconds, 72°C for 30 seconds, 72°C for 5 minutes, hold at 10°C.

#### 3.2.4.1.5 First PCR clean up

The follows the same protocol for sample clean-up in section 3.2.4.1.3 above. The only differences are the following: starting sample volume is 50 ul (instead of 65 ul) and 90 ul SPB are added to each well (instead of 65 ul).

#### 3.2.4.1.6 Quality check using the Bioanalyzer

1 ul of each sample is bioanalyzed using the Agilent DNA 1000 chip described in the Truseq protocol section 3.2.3.1.3. A successful sample tagmentation will reveal DNA fragments ranging from 150-1000bp in size.

#### 3.2.4.1.7 First Hybridization

The purpose of this step is to facilitate the binding of each DNA library to biotinylated oligos (bates). In preparation for hybridization, each sample must be pooled into a library of 12 samples. This requires quantification by the Quibit fluorimetric quantitation system to obtain approximately equal amounts (500ng) of each sample to make a well-balanced library. The Speedvac concentrator and RSB may be used to obtain a volume of 40 ul per sample. This allows for the pooling of samples (in the subsequent step) into a single library, and the ability to run 12X the number of samples per lane, which is both extremely time and cost effective. Next, 40 ul of each library pool is mixed with the following: 50 ul Enrichment Hybridization Buffer, 10 ul Coding Exome Oligos, making at total of 100 ul per sample. The plate is placed on the microshaker followed by centrifugation under standard conditions. The samples are then ready for the NRC HYB program on the thermal cycler under the following conditions: Pre-heat lid to 100°C, 95°C for 10 minutes, 18 cycles of 1 minute incubations, starting at 94°C, then decreasing 2°C per cycle, 58°C for >90 minutes but <24hours.

#### 3.2.4.1.8 First Capture

Capturing post-hybridization requires the use of SPB, which are used to separate genomic DNA-bait hybrids by binding to the biotinylated probes. With a starting volume of 100 ul from the previous step, each library is transferred to a deep well MIDI plate for the SBP cleaning process. 250 ul SPB are added to each sample followed by 5 minutes on the microplate shaker at a slower speed of 1200 rpm. The plate is then incubated at room temperature for 25 minutes followed by centrifugation at standard conditions. The plate is then placed on the magnetic stand for 2 minutes or until the liquid appears clear. Next, all supernatant is removed and discarded without disturbing the beads. After removal from the magnetic stand, 200 ul of Enrichment Wash Solution (EWS) are added to each well, followed by 4 minutes on the microplate shaker at standard conditions. The plate is then incubated on the thermal cycler at 50°C for 30 minutes. After incubation it is once again placed on the magnetic stand, followed by the standard 2-minute waiting period before removing and discarding all supernatant. This step (starting with the addition of 200ul EWS) is repeated a second time for increased purification of target regions.

#### 3.2.4.1.9 First Elution

The following reagents are added to form a pre-mix in preparation for the first elution: 28.5 ul Enrichment Elution Buffer 1 (EEB1) and 1.5 ul 2N NaOH, making 30 ul in total for each pool. 23 ul of pre-mix is added to each well (pool) followed by placement on the microshaker under standard conditions for 2 minutes. This is followed by incubation at room temperature for 2 minutes and then centrifugation at standard conditions. The plate is returned to the magnetic stand and once the liquid turns clear,

21ul supernatant is added to 4 ul Elute Target Buffer 2 (ETB2) followed by microshaking and centrifugation under standard conditions.

#### 3.2.4.1.10 Second Hybridization

A second hybridization to further amplify the DNA and ensure high specificity of the capture regions is required, which occurs between a minimum of 14.5 hours and a maximum of 24 hours. The process is analogous to the first hybridization, but does not require library pooling as this has already been done. Secondly, since the starting volume of each pool is 25 ul after the first elution (and thus for the second hybridization), 15ul RSB are added to the final solution to make 100 ul in total. For more details refer to the section 3.2.4.1.7.

#### 3.2.4.1.11 Second Capture

Once again, samples are thoroughly captured in an identical manner to the process following the first hybridization. Please refer to section 3.2.4.1.8 for more details.

#### 3.2.4.1.12 Capture sample clean up

After the second capture, samples must be cleaned before final enrichment. This process uses 45 ul SPB and is otherwise the same as the “Clean up” in sections 3.2.4.1.3 and 3.2.4.1.5.

#### 3.2.4.1.13 Second PCR Amplification

Finally, an additional PCR amplification step is performed to maximally enrich the library prior to clustering. This requires the addition of 20 ul Nextera Enrichment Amplification Mix (NEM) and 5 ul PCR Primer Cocktail to the 25 ul of each pool. The plate is placed on/in the microshaker and centrifuge under standard conditions and then

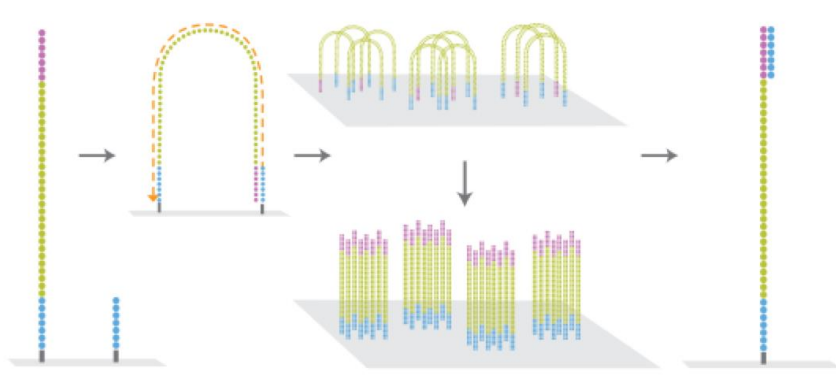
placed on the thermal cycler under the NEM AMP10 program: pre-heat lid to 100°C, 98°C for 30 seconds, 10 or 12 cycles of: 98°C for 10 seconds, 60°C for 30 seconds, 72°C for 30 seconds, 72°C for 5 minutes, hold at 10°C.

#### 3.2.4.1.14 Second PCR clean up

Samples are cleaned in the same manner as section 3.2.4.1.3 (using 90 ul SPB) and then quantified on the bioanalyzer with an Agilent DNA High Sensitivity chip prior to the clustering phase.

#### 3.2.4.2 *DNA amplification and clustering on the C-Bot*

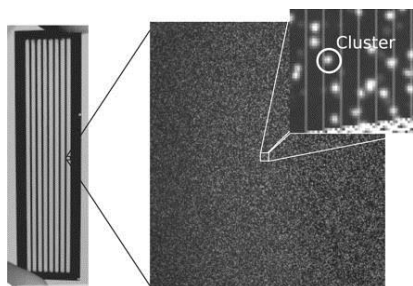
In the second step of the WES protocol, each pool, which consists of 12 libraries, is run on a single lane within the 8-channels located inside each flow cell. A flow cell is sealed glass microfabricated device that allows for cluster generation using an automatic cluster generator (C-Bot). The cluster generation process is initiated by the enzyme DNA polymerase, which amplifies DNA fragments through bridge formation, ultimately producing millions of DNA clusters (Figure 23). Within each distinct cluster, there are roughly 1 million copies of the original fragment, which is required for signal fluorescence and detection during the high throughput sequencing process on the Illumina Hi Seq 2000 (Figure 24).



**Figure 23: C-Bot clustering showing bridge amplification and DNA cluster formation.**

Each unique strand of cDNA is isothermally extended and amplified by DNA polymerase into several hundred million clusters.

(Reproduced from Illumina: [www.illumina.com/documents/products/datasheets/datasheet\\_cbot.pdf](http://www.illumina.com/documents/products/datasheets/datasheet_cbot.pdf)).



**Figure 24: Flowcell with amplified DNA clusters.**

Each cluster contains consists of approximately 1000 identical copies of the unique template. Flow cells facilitate high stability of surface-bound template in conjunction with non-specific binding of fluorescently-labeled nucleotides, allowing bound DNA to interact with key enzymes for amplification.

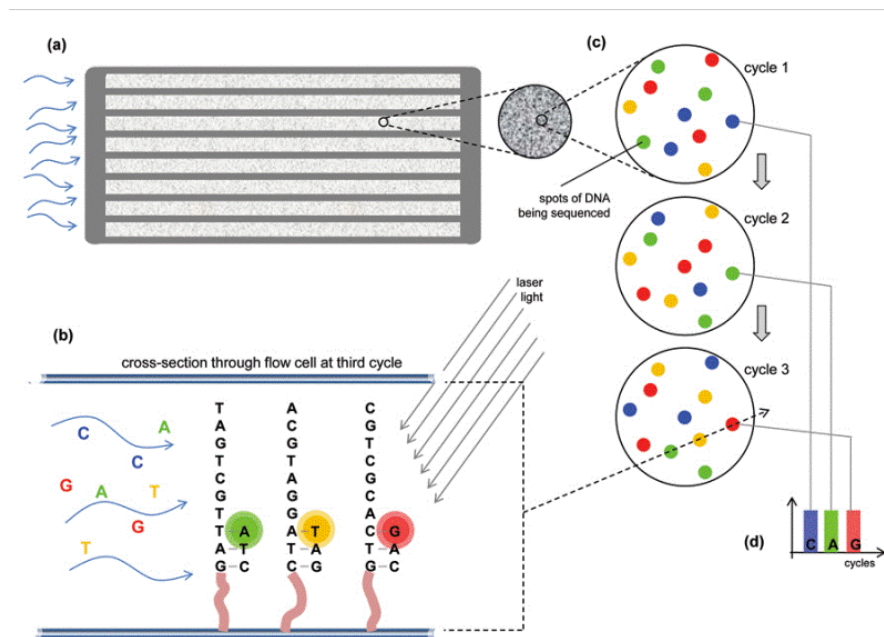
Reproduced from (Whiteford et al, 2009).<sup>337</sup>

### 3.2.4.3 *Parallel sequencing by synthesis on the Illumina Hi Seq 2000*

The third step of WES comprises massive parallel sequencing by synthesis on the Illumina Hi Seq 2000. All four nucleotides are fluorescently labeled with a unique color corresponding to their respective base and are to be incorporated into the oligo-primed cluster fragments on the FC. Linearization of the DNA is accomplished through the cleavage of a single adaptor followed by denaturation to yield single stranded DNA. Next, sequencing primers are added in combination with four reversible terminators,



unique to each nucleotide. Upon the addition of a new nucleotide via DNA polymerase, each base has a 3' hydroxyl group (-OH) chemically blocked. This allows for the optic lens to capture an image after every fluorescently labeled addition, with each FC lane imaged in three distinct 100-tile segments at an approximate cluster density of 30,000 clusters per tile. Once the image has been captured, the 3' blocking group on the hydroxyl group is chemically removed which allows for incorporation of the subsequent base. This process is repeated for approximately 200 cycles in total with read lengths in the 50-100 base pair range. This sequencing process takes place on both single strands of DNA, creating paired-end reads to facilitate accurate mapping during data analysis. The overall run time on the Illumina Hi Seq 2000 consists of approximately 10 days.



**Figure 25: Parallel sequencing by synthesis on the Illumina Hi Seq 2000.**

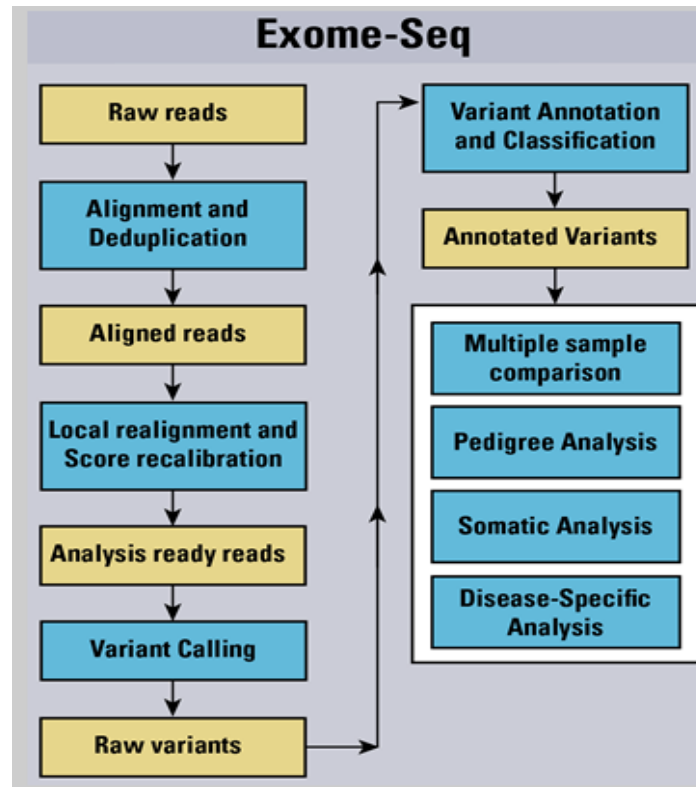
Fluorescently-labeled nucleotides, with a unique color corresponding to their respective base, are added one-by-one into the oligo-primed cluster fragment; this enables the optic lens to capture an image after every fluorescently-labeled addition.

(Reproduced from <http://tucf-genomics.tufts.edu/home/ordering>).

### 3.2.5 Raw data analysis

#### 3.2.5.1 *Mapping, alignment and duplicate removal*

In the fourth step of WES, terabytes of data are transferred from the Illumina Hi Seq 2000 computer to begin the analysis process. First, raw reads of data in the form of fastq files are mapped to their respective genomic origin and appropriately aligned. This is a complex demultiplexing process that is accomplished using Illumina's CASAVA tool and Novoalign's human genome reference (Novocraft technologies). Since fastq files come in pairs in paired-end sequencing, each sample has a forward and reverse sequence for each read. The Phredd score, which is a 10 multiplied by the negative logarithm of the probability of an incorrect base, is used to estimate the confidence in base calling accuracy. Once this is complete, molecular duplicates are excluded and initial alignments are modified via Picard tools (<http://www.picard.sourceforge.net>) (Figure 26).



**Figure 26: Whole exome sequencing analysis pipeline.**

Raw reads generated from the Hi Seq 2000 undergo a series of steps including alignment, de-duplication, local realignment and score recalibration prior to analysis. The next set of steps involves variant calling (including raw variants), annotation and classification tailored to the analytical approach specific to the study.

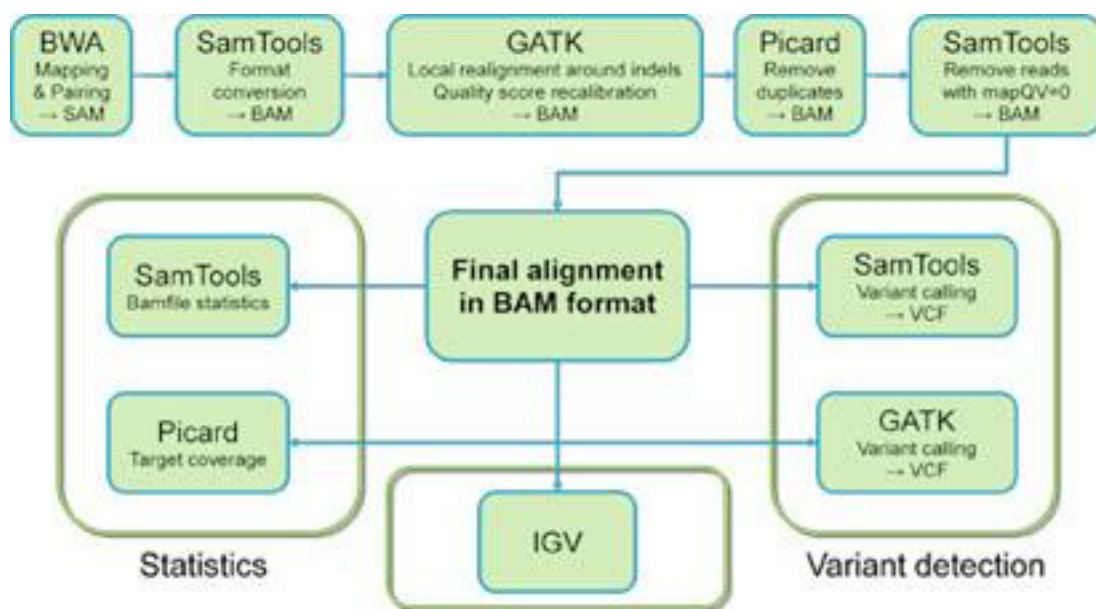
(Reproduced from <http://www.ccmb.med.umich.edu/node/1205>).

### **3.2.5.2 Raw variant callings and file conversions**

In the next phase, fastq files are converted to SAM and BAM files, in which the latter are a compressed and binary version of the former (Figure 27). These files portray human readable mapped sequences with reference sequence coordinates and Phredd scores. Furthermore, BAM and SAM files allow one to analyze all sites with variant calls (some real, other artifacts) in the form of SNPs and indels. BAM files can also be visualized using the computer program, Interactive Genomics Viewer (IGV) to determine

the exact number and quality of reads called in favor of the wildtype and alternate alleles.

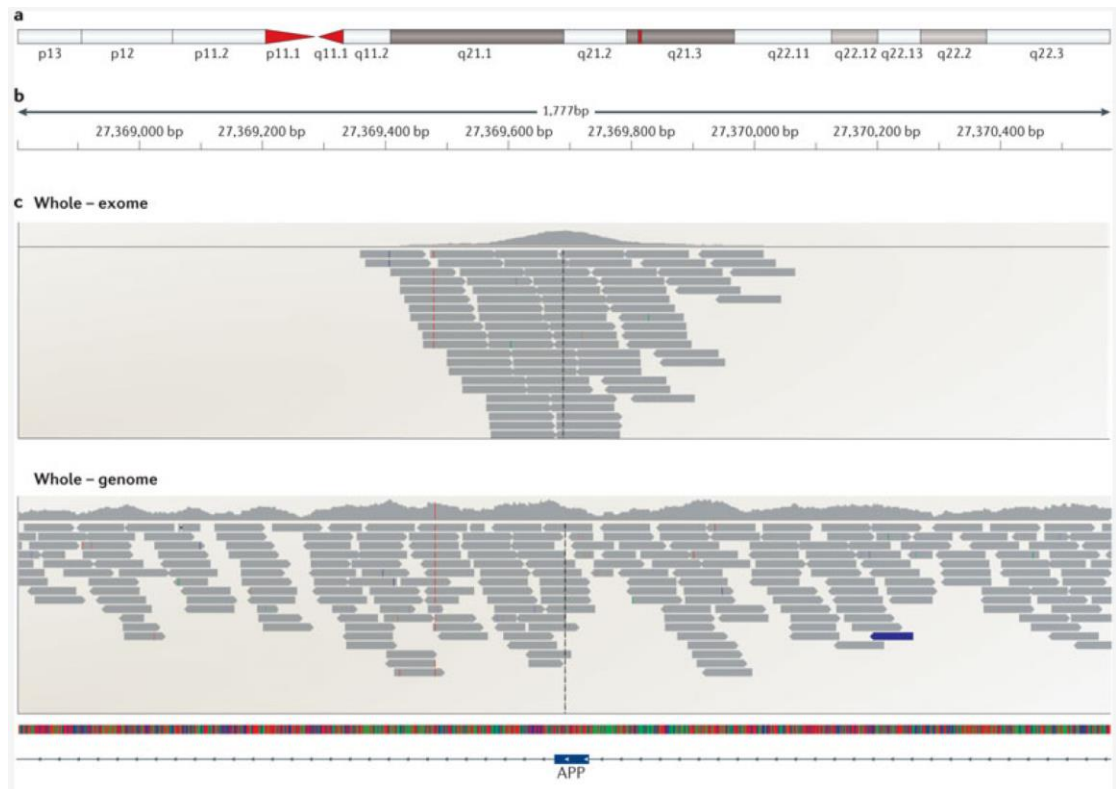
An example of this is shown in Figure 28.



**Figure 27: Bioinformatics Pipeline of Raw Variant Callings and File Conversions.**

Samples undergo a series of processing steps after completion of sequencing by parallel synthesis on the Hi Seq 2000. Primarily this includes mapping, pairing, and format conversion to create preliminary BAM files. BAM files must undergo further manipulation including local realignment around indels, quality score recalibration, duplicate removal, and elimination of low quality reads to generate analysis-ready BAMs. Subsequently, these files can be viewed on IGV, with a focus on variants determined by the GATK generated VCF and statistical analyses attained via SamTools and Picard.

Reproduced from (<http://www.ikmb.uni-kiel.de/research/genetics-bioinformatics/genome-exome-analysis>).



**Figure 28: Read coverage of *APP* using whole exome sequencing and whole genome sequencing.**

Reads can be visualized as shown in Interactive Genome Viewer using BAM files. The top image demonstrates reads obtained from WES, which only covers coding regions. The bottom image depicts both intronic and exonic regions of *APP* generated by WGS. *APP* = Amyloid Precursor Protein.

(Reproduced from Bras et al 2012). <sup>7</sup>

### 3.2.5.3 Incorporation of reference databases

In a very computationally intensive process, extensive integration merges relevant information about known variation sties, LD, family structure, population substructure with raw variant calls.

#### **3.2.5.4 Assignment of quality scores (Phredd scores) to all variant calls**

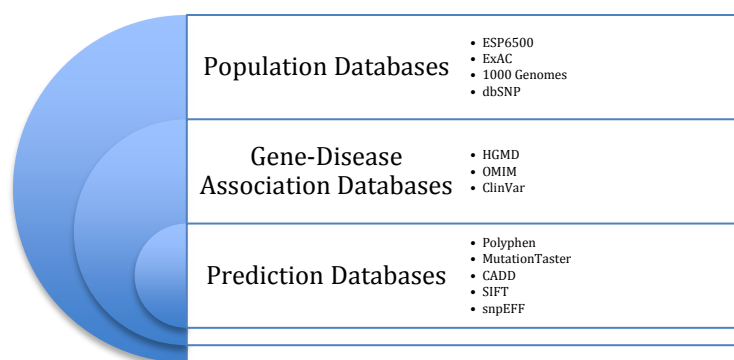
Each variant call is associated a Phredd score to denote the level of confidence in accuracy. Based on the Phredd Score, the most probable genotype is determined for every sample at every site of variation using the Genome Analysis tool Kit (GATK, <http://www.broadinstitute.org/gatk/>).

#### **3.2.5.5 Generation of variant call files (VCF) and (group) gVCFs**

Finally, samples can be subsetted at the user's discretion to form tab-delimited Variant Calling Files (VCF) to be extensively annotated using Annovar (<http://www.openbioinformatics.org/annovar>), VCFtools (<http://www.vcftools.sourceforge.net/>) and PLINK (<http://pngu.mgh.harvard.edu/~purcell/plink/>). VCFs include individual genotypes and variant calls for each sample, including both SNPs and indels.

#### **3.2.5.6 Downstream analysis and filtering of VCFs**

Using Annovar, several annotations can be performed on the VCF to manipulate it as desired. This allows exclusion of common variants using frequencies from public databases (dbSNP, 1000genomes, ESP6500), filtering based on MAF, and affords the ability to predict deleterious variants using web-based available programs (Polyphen2, SnpEFF, CADD, MutationTaster). A Table of Public Databases used for filtering analysis is listed below (Figure 29).



**Figure 29: Public Reference Databases used as exclusion criteria for WES analysis.**

A combination of population, gene-disease association and predication databases are used for variant filtering in ANNOVAR.

### 3.2.6 Looking for variants in PD risk and causal genes

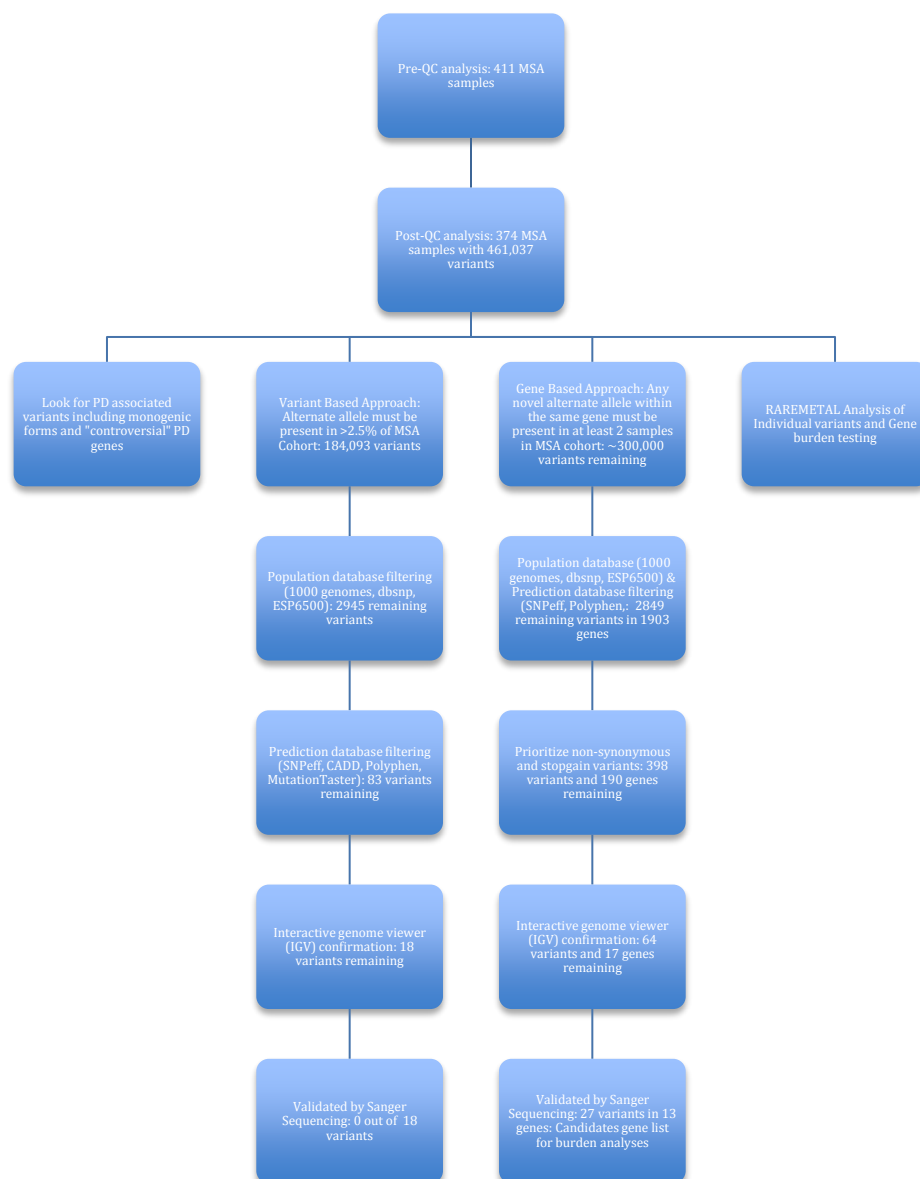
As there are no known risk variants or genes for MSA, we thought it was important to look for those associated with alpha-synucleinopathies, most prominently PD. Further, given the estimated 14% likelihood of misdiagnosis among clinically diagnosed MSA samples, this was a critical step to eliminate any true PD cases from our cohort, an important part of the association analysis.<sup>322</sup> The genes investigated included several categories of PD associated genes: first, all genes harboring causal variants attributed to monogenic forms of PD, described in detail in section 1.4.2. Second, we also incorporated all PD associated genes that have been deemed controversial, as the results supporting their significance consistently fail independent replication in other cohorts.

### 3.2.7 Variant and Gene based Approach Filtering Pipelines

Downstream filtering is tailored to the type of study (sporadic or familial), suspected mode of inheritance (autosomal dominant, autosomal recessive, x-linked, mitochondrial, *de novo*) and level of penetrance (complete, incomplete). With MSA being a predominantly sporadic disease we initially used less stringent filters. The variant

filtering pipeline used for exome analysis is demonstrated in Figure 30. The first step performed was QC analysis to remove samples that did not meet our pre-determined QC standards. With several hundred thousand variants remaining, we focused on 4 unique analyses to obtain a more manageable number of candidate genes and variants. First we looked at PD associated variants in both monogenic forms as well as more controversial PD genes. Secondly, we used a variant-based approach to search for very rare shared variants present in at least 2.5% of the MSA cohort. Thirdly, we incorporated a gene-based approach to identify all novel alternate alleles within the same gene in at least 2 MSA samples. Finally, we performed individual variant and gene burden analyses using RAREMETAL, which will be discussed in detail in section 3.2.10.





**Figure 30: Variant Filtering Pipeline used for MSA whole exome sequencing analysis.**

Four primary analyses were incorporated in this workflow: investigation of PD associated variants, a Variant Based Approach, a Gene Based Approach, and RAREMETAL analyses.

### 3.2.8 Annovar Filtering Process

Using annovar, we used several commands on terminal to execute each filtering step. The first consisted of converting the VCF to an annovar “readable” format (perl

convert2annovar.pl). The read-out for this included the number of specific types of variants: homozygotes, heterozygotes, SNPs, and Indels; further, among the SNPs it specified the precise number of transitions and transversions. Next, we used the – geneanno command to create two output files: an all variant file and an exonic file for preliminary annotation. From the exonic file, we filtered for the most damaging mutations using the “awk” program functions and reorganized columns into their appropriate location in the VCF. These changes included: stop gain, stop loss, nonsynonymous, frameshift and splicing. Upon downloading all of the reference genome databases from annovar, including 1000genomes, ESP, dbsnp138, and several prediction sites, among others, we applied these filters to obtain preliminary lists of candidate genes. To further manipulate our candidate list, we applied MAF filters (i.e.  $MAF < 0.01$ ) to remove all variants with a MAF above a defined threshold. Finally, we used awk commands to create separate homozygous, heterozygous and compound heterozygous files for further analysis and exploration on IGV.

### **3.2.9 Sanger sequencing confirmation of variants**

After verifying that a particular variant appears real on IGV, it is important to confirm the presence and genotype of this variant through an independent method, most often Sanger sequencing. The protocol for Sanger sequencing is as follows:

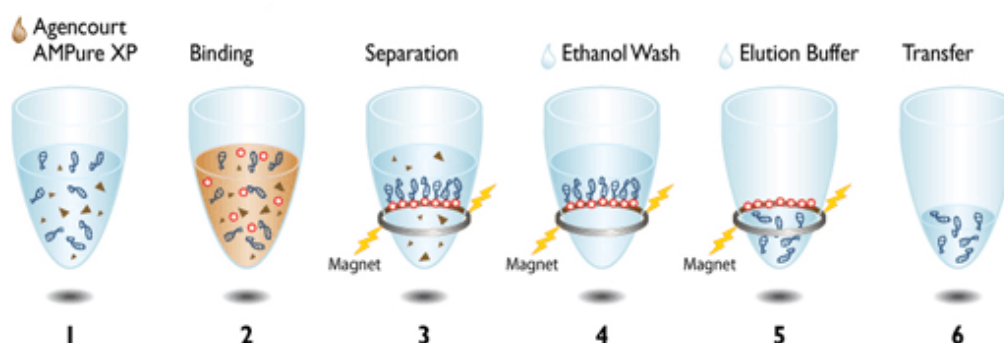
Primers must be designed to cover the variant using Primer3 (<http://bioinfo.ut.ee/primer3-0.4.0/>) and the University of California Santa Cruz (UCSC) genome browser (<http://genome.ucsc.edu/>) reference sequences. Oligocalc (<http://biotools.nubic.northwestern.edu/OligoCalc.html>) can be used a quality control measure to check for hairpin turns or single primer self-dimerization. All primers used for

this study are listed in the Appendix section 8.1.2.

Primers must be then be optimized to the appropriate thermal cycler PCR settings once a working solution is made. This often requires some troubleshooting due to variation in melting temperatures ( $T_m$ ), GC content and several other factors. All thermal cycler conditions are listed in the Appendix section 8.1.3 Assessment of a specific and efficient PCR amplification is performed by visualization and appropriate sizing of the product on an agarose gel containing ethidium bromide.

Once primers are optimized, a working solution is made specific to the stock concentration received. The initial PCR reaction requires 12 ul FastStart PCR Mastermix (Roche, IN, USA), 1 ul Forward primer, 1 ul Reverse primer, and 1 ul gDNA (around 10ng/ul). This reaction is mixed and centrifuged and placed on the thermal cycler at the appropriate settings determined during optimization. This usually takes between 1.5-2.5h, depending on the number and duration of cycles, and can followed by a cleanup on the Biomek FX robot (Beckman Coulter, CA, USA) in order to remove un incorporated dNTPs, primers, salts and DNA polymerase using Ampure magnetic beads (Agencourt Bioscience Corporation, MA, USA). 27 ul of paramagnetic beads are added to each sample (15ul) followed by thorough mixing and incubation for 5 minutes at room temperature, allowing the paramagnetic beads to bind the amplified PCR fragments. The plate is then moved to a magnetic Agencourt SPRIplate to separate the beads from the solution, followed by the aspiration and discarding of the supernatant solution consisting of dNTPs and unbound primers. A series of two washing steps using a 70% ethanol solution is performed. Next, the cleaned PCR amplicons are resuspended in 30 ul distilled and deionised molecular grade water and transferred into a separate 96 well

PCR plate for further sample processing. A figure illustrating the cleaning process is shown below in (Figure 31).



**Figure 31: Sanger sequencing PCR cleanup using Ampure paramagnetic beads.**

Paramagnetic beads bind the amplified PCR fragments and are separated from the solution containing dNTPs and unbound primers upon binding to the magnet. Two subsequent washes with ethanol followed by resuspension in elution buffer creates a clean sample ready for Sanger sequencing.

Reproduced from ([http://www.agencourt.com/products/spri\\_reagents/ampure/](http://www.agencourt.com/products/spri_reagents/ampure/))

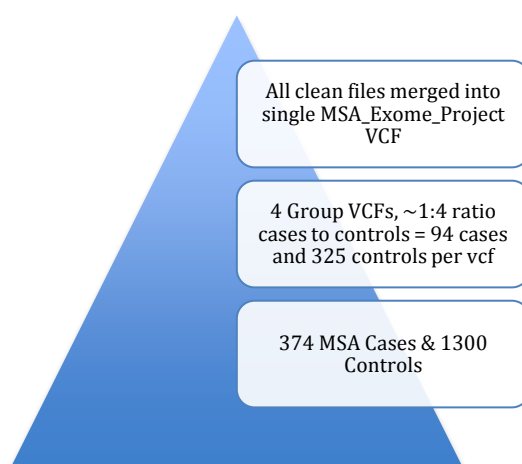
After PCR amplification product purification, the plate is prepared for bi-directional direct dye-terminator sequencing using the BigDye chemistry (v.3.1, Applied Biosystems, CA, USA). The sequencing reaction recipe is the following: 5 ul cleaned PCR product, 0.5 ul BigDye (v.3.1), 1 ul of 10nM primer (forward or reverse), 1.875 ul Sequencing Buffer (Applied Biosystems), 1.875 ul deionized molecular grade water (Mediatech. Inc., VA, USA). The plate is adequately mixed and centrifuged and ready for sequencing on the thermal cycler block.

To remove excess fluorescent dye-terminator and other contaminants, CleanSEQ paramagnetic beads (Agencourt Bioscience Corporation, MA, USA) are used according to the manufacturer's protocol. This cleaning requires the addition of 10 ul CleanSEQ and 45 ul 85% ethanol solution to the 10 ul sequencing reaction form above. Thorough mixing allows the paramagnetic beads to bind the sequenced amplicons. Next, the plate is

moved to the magnetic Agencourt SPRI Plate for 3 minutes in order to separate the bead-bound sequencing amplicons from any contaminants. After this incubation, the supernatant solution is aspirated and discarded, followed by an additional round of washing with 85% ethanol. Lastly, the cleaned sequencing products are eluted from the paramagnetic beads in 40ul of distilled and deionized molecular grade water and transferred to a clean 96-well semi-skirted reaction plate for further processing. Septa sealing is placed on semi-skirted plates with purified sequences and are analyzed on the ABI 3730 DNA Analyzer (Applied Biosystems, CA, USA). Electropherograms are visualized using Sequencer software (version 4.2 Gene Codes Corporation, MI, USA).

### 3.2.10 Individual Variant and Gene Burden Testing

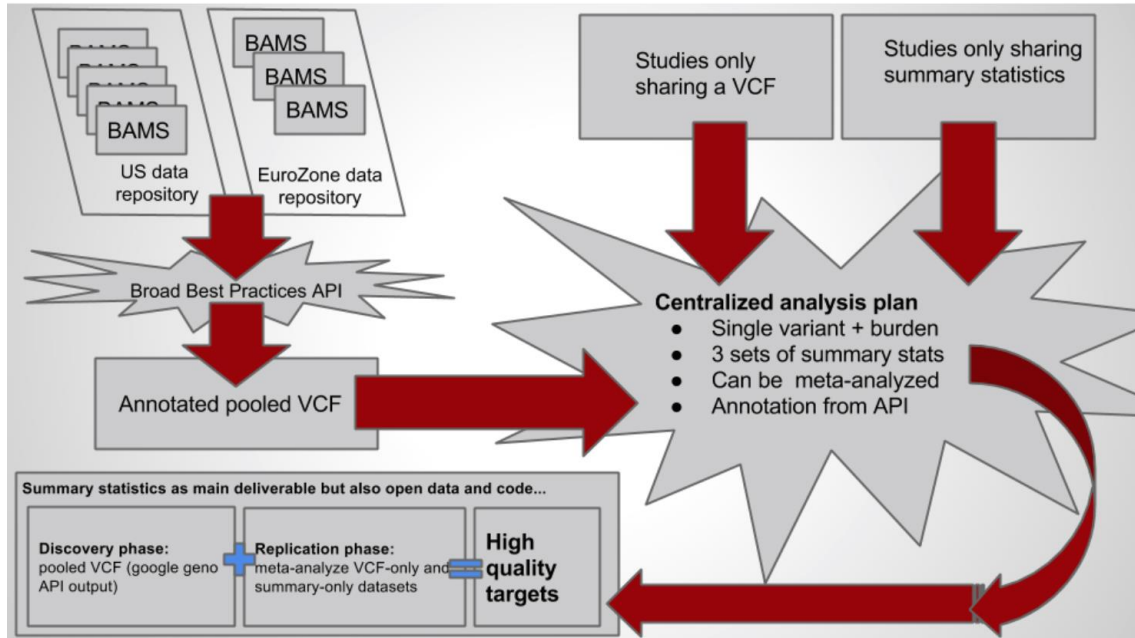
By analyzing the clean MSA cohort (374 samples) with approximately 4X the number of controls, our goal was to obtain sufficient statistical power upon performing individual variant and gene burden tests (Figure 32).



**Figure 32: Merging clean MSA sample cohort with 1300 control samples to make final VCF**

To create a VCF ready for analysis, 4 group VCFs (1 case, 3 controls) were merged.

The second phase of analysis rested on an association based analysis rather than a filtering based approach. Because an association-based approach is extremely sensitive to differences in the way in which cases and controls are analyzed we pursued a single unified reanalysis approach. This approach, which used exome sequencing data from 374 MSA cases and 1300 controls, was made possible by the use of GoogleGenomics within GoogleCloud. Thus WES data from cases and controls were recalled and aligned together on GoogleCloud according to the pipeline shown below (Figure 33). Fundamentally, using the cloud allows for recalling and realignment of large-scale WES project data based on the latest versions of dbGaP. This enhances the sensitivity of rare variant detection, some of which may be missed when only using local alignment tools in the LNG. Our approach used BAMs of cases and as well as controls from US and European datasets to generate an annotated and pooled VCF on our local drive. This is the VCF utilized for WES variant and gene based association pipelines described in sections 3.3.3-3.3.4. On GoogleCloud, data is stored through a “bucketing” method, followed by reprocessing to increase data quality and alignment parameters. This data is then transferred back to the local drives and can be analyzed using RAREMETAL to generate summary statistics from several gene burden and single variant tests. As this is the primary phase of analysis, we consider the results, while valuable and interesting, as hypothesis generating or part of a discovery phase. Any significant findings will need to be followed-up in a replication phase.



**Figure 33: Pipeline used for data pooling on googleCloud before running RAREMETAL analyses.**

BAM files from the US and EuroZone data repositories were combined using API to form a single annotated pooled VCF cohort. Using RAREMETAL analysis, single variant and gene burden analyses were performed on the group VCF file to initiate the discovery phase of hypothesis generating results. API = Application Program Interface.

(Reproduced courtesy of Dr. M Nalls.)

### 3.2.10.1 RAREMETAL Analysis: Quality Control

To check for population stratification between cases and controls, the first two PCA covariates were plotted as eigenvectors, which represent the vast majority of ethnic variability among all cases and controls. The next step involved the generation of Quantile-quantile (QQ) plots, which is an important quality control parameter, as it represents a two-dimensional plot of the chi-squared test comparing the observed and expected p-values. A chi-squared test is performed for all markers included in the study and the p-values are plotted as  $-\log_{10}(\text{observed p value})$  along the y-axis. The x-axis plots the  $-\log_{10}(\text{expected p value})$  and any deviation from  $x=y$  line suggests one of the following: first, if there is population stratification, deviation would be expected along

the full length of the  $x=y$  line. Second, if there is true association, the line will deviate only towards the latter half of the  $x=y$  line.<sup>36,338</sup> It is important to note that plotting this will not allow one to adjust data for population stratification without running a prior PCA to determine (and remove) ethnic outliers.

### 3.2.10.2 RAREMETAL analysis: gene burden and single variant testing

This requires use of the command-line and allows one to input a VCF or PED file with genotypes and utilize this information to create summary statistics through several different tests. It can be applied to familial or sporadic analyses, and generates visual plots revealing both quality control statistics (i.e. QQ plot), as well as a Manhattan plot with the most significant results. Further, RAREMETAL has the ability to incorporate variance-covariance matrices to implement conditional analyses, which can differentiate true signals from those derived by nearby variants. The program utilizes 4 different types of tests to calculate individual variant and gene burden analyses, each with unique characteristics to obtain the most comprehensive and statistically significant set of results. At the meta-analysis stage, RAREMETAL allows one to organize variant groups based on gene-level statistics, creating unique reports for every gene-level test (<http://genome.sph.umich.edu/wiki/RAREMETALWORKER>). In essence, this illustrates the fundamental approach behind RAREMETAL meta-analysis, which is the extrapolation of single variant score statistics (calculated using the Cochran-Mantel-Haenszel method) with known LD relationships into gene-level test statistics with corresponding p-values to reflect significance.



### 3.2.10.2.1 Single variant analysis: variance component (non-burden) tests

#### 3.2.10.2.1.1 The SKAT

The SKAT aggregates associations between variants and phenotype using a kernel matrix defined by SNP-SNP interactions. As a non-burden test, the SKAT is more powerful when a substantial fraction of variants are non-causal or the effects of associated variants are in different directions.<sup>339</sup> Further, the SKAT can also apply covariates, thus allowing for the incorporation of continuous traits. Notably, the SKAT is considered a particularly sensitive test that allows one to detect both protective and risk variants associated with disease. A caveat to this, however, is that the SKAT is less powerful than a burden test if the majority of rare variants are truly causal and in the same direction. Further, variance component tests are generally not the most stable for small cohorts with significantly different number of cases and controls.<sup>340</sup>

### 3.2.10.2.2 Weighted Aggregation Test

#### 3.2.10.2.2.1 *The Madsen-Browning (MB) burden test*

As a standard burden test, the Madsen-Browning burden test requires that all rare variants are collapsed within a single gene into a single burden variable. Consequently, this measures the cumulative effects of rare variants in a gene through regressions between burden variable and the gene of interest, adhering to a multivariate normal distribution.<sup>340</sup> This test is considered a nonparametric weighted sum test (WST) by assigning a unique "weight" to every variant site; consequently, these weights are merged to determine the overall aggregated burden (<http://varianttools.sourceforge.net/Association/Weighted>). As a burden test, there is an

implicit assumption that all rare variants within a particular region influence the phenotype in the same direction and with a relatively similar magnitude.<sup>340</sup> However, when these assumptions regarding causality, directionality and magnitude are not upheld, there is a significant loss of power in detection.

### 3.2.10.2.3 Adaptive Burden test

#### 3.2.10.2.3.1 Variant Threshold (VT) test

The VT test is a unique type of burden test in that it selects the ideal allele frequency threshold (via MAF) for all rare variants in a gene to enhance detection of genes with the greatest burden. While this is extremely powerful, the same assumptions of a classic burden test regarding causality, directionality and magnitude apply; hence, when not upheld, there will be a substantial depreciation of power (<http://genome.sph.umich.edu/wiki/Rvtests>).

### 3.2.10.2.4 Combined Burden test

#### 3.2.10.2.4.1 Combined Multivariate and Collapsing (CMC) test

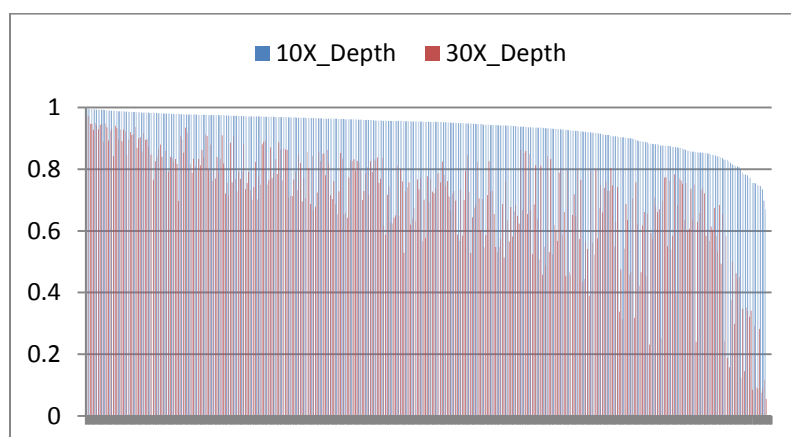
This test is ideal for investigation rare variants within individual genes to assess overall burden of a single gene. All variants within a single gene are considered a test unit and collapsed into a binary system, in which a region is coded as “0” when all rare variants are wildtype and given a “1” if any rare variant harbors the minor allele.<sup>341</sup> The next step involves merging of common variants in the same gene with the coded rare variants to generate a multivariate model, testing the null hypothesis: the absence of an association between the specific gene and disease of interest. By using Fischer’s test to

investigate this hypothesis, an exact p-value is generated, making this a favorable burden test for evaluating statistical significance of single gene burdens.

### 3.3 Results

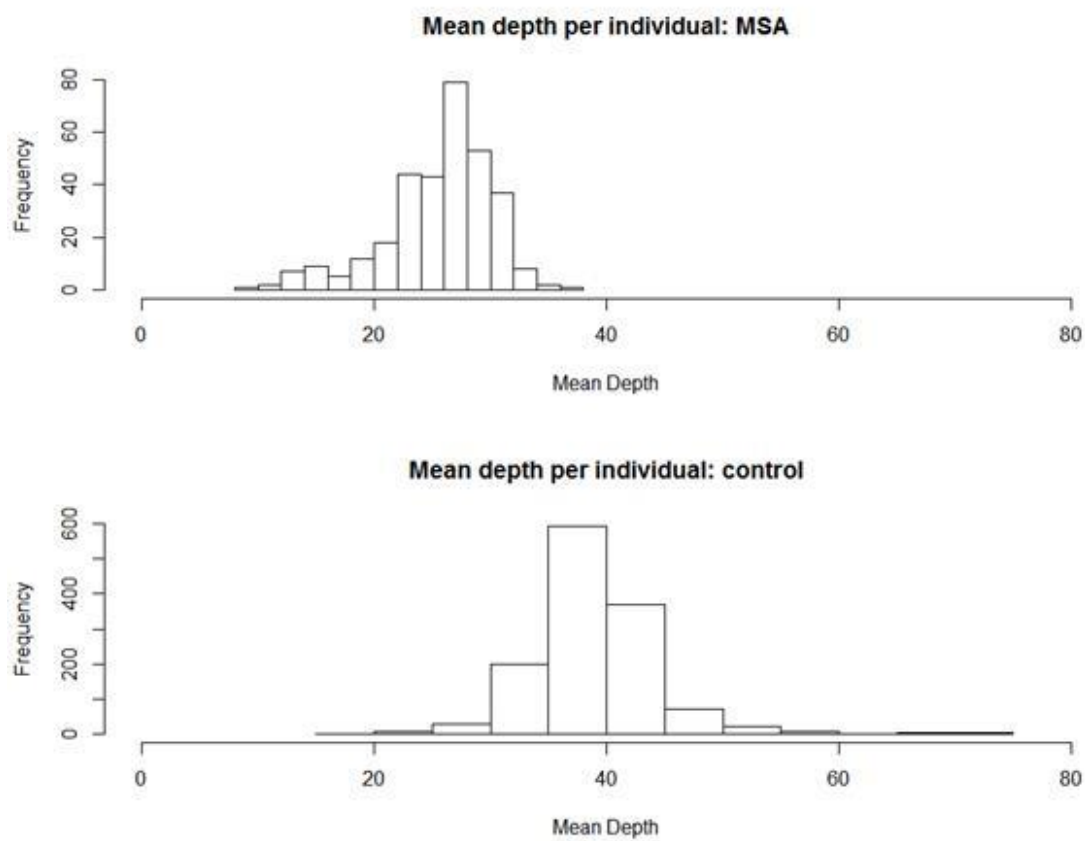
#### 3.3.1 Quality control filtering of locally executed analyses

In the preliminary stages of the variant filtering pipeline using data derived from local analyses, we needed to eliminate samples that did not meet quality control standards. However, by re-preparing or running additional lanes to obtain increased coverage and depth, the final quality attained by MSA samples was very good (Figure 34). The samples that did not meet the following criteria were eliminated from the cohort: >90% 10X depth, >70% 30X depth and a PCR duplicate rate <14%. Further, it was important to compare the depth with the 1300 controls used to prevent any bias (Figure 35).



**Figure 34: 10X and 30X depth of 411 MSA exome sequenced samples.**

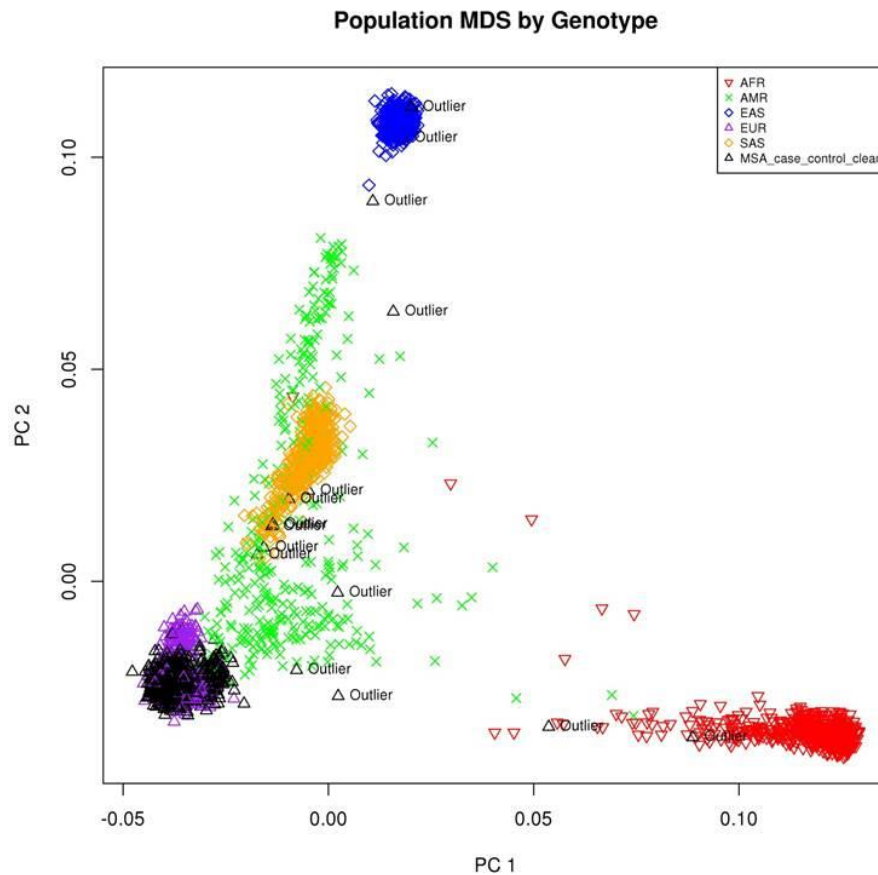
This reflects the combined results of both Illumina Truseq and Nextera Protocols.



**Figure 35: Mean depth per individual in MSA samples (top) and 1300 controls (bottom)**

Mean depth per individual averaged around 39X for controls and 30X for MSA samples, respectively.

In the first stage of the filtering pipeline, quality control measures were necessary to remove the following: call rate and heterozygosity rate outliers, individuals that were related or duplicated, those with missing genders and samples deemed ethnic outliers based upon multidimensional scaling (MDS). A MDS plot of all MSA samples is shown below. Any ethnic outliers were removed from subsequent analyses (Figure 36).



**Figure 36: Multidimensional scaling of 411 MSA samples to identify and remove population outliers**

The first two principal components were used for population stratification. A colored key is located in the upper right hand corner.

After eliminating samples using quality control measures, we were able to proceed with analyzing the “clean” MSA cohort consisting of 374 samples.

### 3.3.2 Looking for variants in PD associated genes

Among the “clean” case cohort, a total of 11 variants were found in 8 PD associated genes, though notably most are not monogenic (Table 11). Among these, only a single heterozygous variant was revealed in *LRRK2*. Despite the dubious appearance of the BAM file at this location, we followed this up with Sanger sequencing. Sanger

sequencing results revealed this was indeed an artifact. With no other putative PD causal variants in the locally processed case/control cohort, we proceeded with our analyses.

CHR	POSITION	ID	REF	ALT	GENE	EFFECT	SANGER CONFIRMED	#, Cases het, hom	# in Controls het, hom	Cases % het, hom	Controls % het hom
2	233708909	.	AAGC	AAGCAGC,A	GIGYF2	nonframeshift insertion	X	1,0	0,0	0.26%	0.00%
2	233712227	rs66736190	ACAG	*,AGCAGCAGG	GIGYF2	frameshift substitution	X	11,0	40,0	2.90%	3%
3	184039743	.	GGAA	G	EIF4G1	nonframeshift deletion	X	3,0	9,0	0.80%	0.67%
8	16859372	.	AGTCTGTG		FGF20	frameshift insertion	NO	4,0	0,0	1.10%	0%
12	40749989	.	C		LRRK2	frameshift insertion	NO	1,0	0,0	0.26%	0%

**Table 11: Results of variants among all MSA samples in PD associated genes.**

Total number of cases: 374. Total number of controls: 1324. In the “sanger confirmation” column, variants with a “NO” were confirmed to be artifacts. All other variants that were either present in controls or did not have damaging predictions were not followed up with sanger sequencing, indicated by an “X”. A variant in *LRRK2* is highlighted in red as it is the only gene identified to harbor variants causal for monogenic forms of familial PD.

Starting with 461,037 variants, we focused on two approaches for analysis: a variant based approach and a gene based approach (Figure 30).

### 3.3.3 Filtering through a Variant Based Approach

Our rationale towards using a variant based approach was based the idea that very rare and potentially novel variants that are shared by a certain percentage of the full case cohort suggests a plausible association. As there are currently no known genes associated with MSA etiology, we were completely unbiased regarding our initial candidate gene list. By using the criteria that the alternative allele must be present among >2.5% of the full MSA case cohort, we were left with 184,093 variants (Figure 30). Next, we filtered using population databases including those listed in Figure 29 (1000 genomes, dbsnp, ESP6500), bringing us down to only 2945 variants. To further distill our candidate list, we used several strong prediction filters, including SnpEff, Polyphen and one of the most stringent, CADD, to obtain a list of 83 variants predicted to be very damaging. As this number is feasible to work with, all candidate variants were scrutinized on IGV to assess

the number of reads, direction of reads, mapping quality, and Phredd score of each sample allegedly carrying the variant according to the annotated VCF. Notably, certain genes are notorious for false positives, such as variant calls in “*SARM*” and “*HRNR*”; to remedy this, all cases were visually compared with several control samples previously run in the laboratory. After eliminating several candidates due to sample misalignments and artifacts, the remaining list consisted of 18 heterozygous variants in 18 genes. Unfortunately among the 18 candidates we were unable to confirm these variants via Sanger sequencing (Table 12). While the variant calling procedure has worked extremely well in family based analyses, the high failure to replicate rate led us to re-examine this approach in the context of association analyses. With this in mind the data was reanalyzed using a unified calling approach in GoogleGenome, as described later in this chapter.

	Gene	CHR	POS	ID	REF	ALT	Effect	Sanger confirmed
1	DNAJC11	1	6727802	.	TTC	T	Frameshift	NO
2	PRAMEF2	1	12921594	rs201429745	C	G	Stop_gained	NO
3	FAHD2B	2	97749524	rs187104986	C	A	Stop_gained	NO
4	SLC20A1	2	113416606	.	CAG	C	Frameshift	NO
5	RAPGEF6	5	130815368	.	C	CT	Frameshift	NO
6	RBM27	5	145647319	.	T	TA	Frameshift	NO
7	FOXO3	6	108985176	rs34133353	T	TG	Frameshift	NO
8	ABCB4	7	87074281	.	G	GA	Frameshift	NO
9	VCP	9	35059646	.	A	AT	Frameshift	NO
10	KCNMA1	10	78729785	.	C	CT	Frameshift	NO
11	TYRO3	15	41862897	.	G	GGAGA	Splice site donor	NO
12	CHD2	15	93540315	.	G	GA	Frameshift	NO
13	PHLDA1	12	76424933	.	GC	G	Frameshift	NO
14	IPP	1	46184896	rs144663569	AAC	A	Frameshift	NO
15	PTMA	2	232577559	.	T	C	Stop_lost	NO
16	FAM104B	X	55172630	rs113263757	G	A	Stop_gained	NO
17	DNHD1	11	6567913	.	C	CT	Frameshift	NO
18	ADRA2B	2	96780997	.	CAG	C	Frameshift	NO

**Table 12: Sanger sequencing results using variant-based approach for MSA WES analysis**

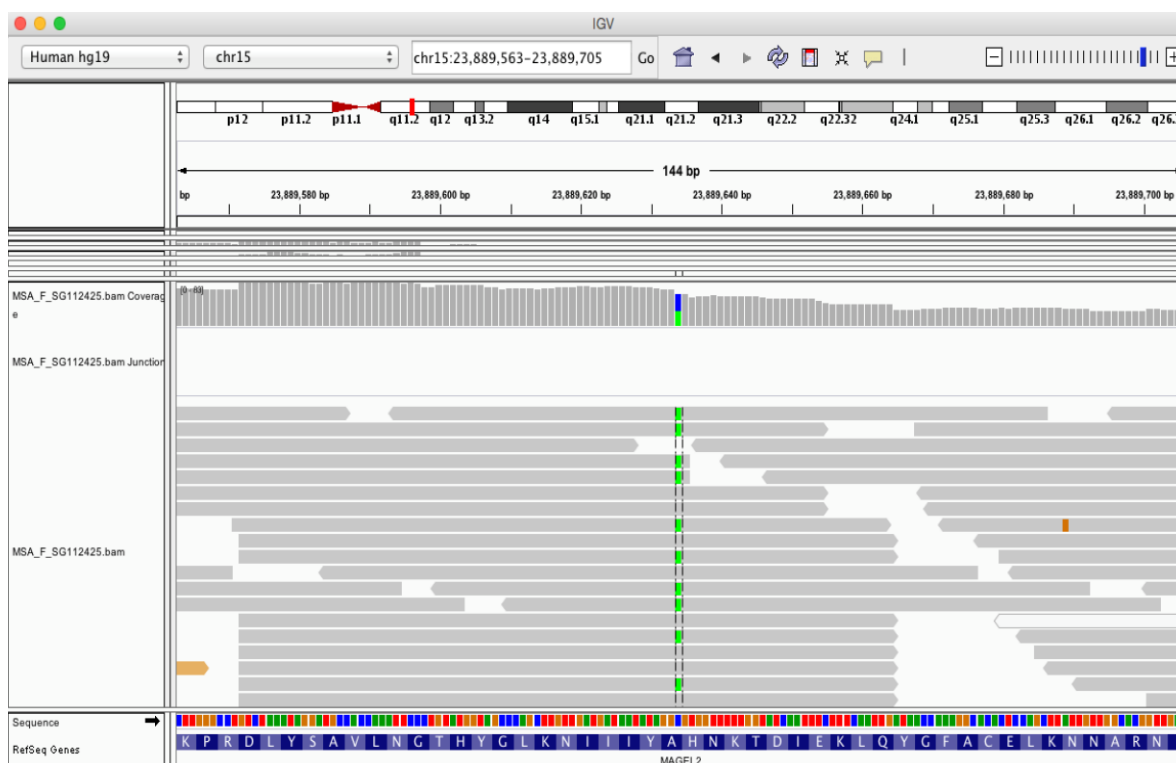
The location, corresponding gene, dbSNP ID, mutation effect and sanger sequencing validation status is listed for each candidate variant.

### 3.3.4 Filtering through a Gene Based Approach

As an alternative approach, we decided to be less restrictive than the former pipeline by using the following criteria: at least 2 individuals in the MSA cohort must carry a novel variant within the same gene. Starting once again from 461,037 variants, we applied population databases filters and fewer prediction filters (not including CADD) to yield a total of 2849 variants in 1903 genes. As this was still a substantially large list, we needed to prioritize by visualizing them in IGV.

Notably, the results of the variant-based gene analysis suggested that a substantial proportion of WES artifacts are indels and frameshifts; alternatively, WES technology is much more suitable for single variant (SNP) detection. Hence, we further subdivided our candidate list into the following two categories: non-synonymous, stop\_gain and stop\_loss SNVs, and everything else that was not a SNV (indels, frameshifts variants). We designated the former group as our primary focus, consisting of 398 variants in 190 unique genes, given the greater likelihood of authentic data quality from SNVs (vs. indels, frameshifts etc). After a comprehensive search through IGV and being particularly rigid in our candidate selection, a total of 64 novel variants in 13 genes were promising (Table 13, Figure 37):





**Figure 37: IGV example of high quality depth and coverage in candidate gene, *MAGEL2***

Two of our candidate variants located in *MAGEL2* demonstrated excellent depth on IGV reads and were later confirmed with sanger sequencing. One of these two variants is illustrated above.

# Samples Clin v Path	#CHR	POS	REF	ALT	GENE NAME	ID	Unique Effect	Sanger Confirmation, Genotype(s)
1, Clin	chr22	18121566	G	A	BCL2L13	.	nonsynonymous SNV	NO, wt/wt
1, Clin	chr22	18121557	C	T	BCL2L13	.	nonsynonymous SNV	YES, wt/mt
1, Clin	chr22	18121587	G	T	BCL2L13	.	nonsynonymous SNV	YES, wt/mt
1, Path	chr6	90577545	G	A	CASP8AP2	.	stopgain SNV	X
1, Path	chr6	90573769	G	A	CASP8AP2	.	nonsynonymous SNV	X
<b>1, Clin</b>	<b>chr6</b>	<b>90573431</b>	<b>T</b>	<b>C</b>	<b>CASP8AP2</b>	.	<b>nonsynonymous SNV</b>	<b>YES, wt/mt</b>
<b>1, Path</b>	<b>chr6</b>	<b>90576144</b>	<b>A</b>	<b>G</b>	<b>CASP8AP2</b>	.	<b>nonsynonymous SNV</b>	<b>X</b>
1, Clin	chr6	90577775	A	G	CASP8AP2	.	nonsynonymous SNV	YES, wt/mt
1, Clin	chr6	90572365	T	G	CASP8AP2	.	nonsynonymous SNV	YES, wt/mt
1, Clin	chr11	11373751	G	A	CSNK2A3	.	stopgain SNV	YES, wt/mt
1, Clin	chr11	11373910	C	A	CSNK2A3	.	nonsynonymous SNV	YES, wt/mt
1, Path	chr3	53346386	C	A	DCPIA	.	nonsynonymous SNV	X

1, Clin	chr3	53322243	G	A	DCP1A	.	nonsynonymous SNV	YES, wt/mt
1, Path	chr3	53326508	T	G	DCP1A	.	nonsynonymous SNV	X
1, Clin	chr3	53376290	G	A	DCP1A	.	nonsynonymous SNV	YES, wt/mt
1, Clin	chr3	53381504	G	A	DCP1A	.	nonsynonymous SNV	NO, wt/wt
<b>1, Clin</b>	<b>chr7</b>	<b>1786631</b>	<b>T</b>	<b>G</b>	<b>ELFN1</b>	.	<b>nonsynonymous SNV</b>	<b>YES, wt/mt</b>
<b>1, Path</b>	<b>chr7</b>	<b>1784707</b>	<b>G</b>	<b>A</b>	<b>ELFN1</b>	.	<b>nonsynonymous SNV</b>	<b>X</b>
<b>1, Clin</b>	<b>chr7</b>	<b>1785179</b>	<b>G</b>	<b>A</b>	<b>ELFN1</b>	.	<b>nonsynonymous SNV</b>	<b>YES, wt/mt</b>
<b>1, Path</b>	<b>chr7</b>	<b>1784309</b>	<b>G</b>	<b>T</b>	<b>ELFN1</b>	.	<b>nonsynonymous SNV</b>	<b>X</b>
<b>2, Clin</b>	<b>chr7</b>	<b>150434680</b>	<b>G</b>	<b>A</b>	<b>GIMAP1-GIMAP5</b>	.	<b>nonsynonymous SNV</b>	<b>YES, both wt/mt</b>
1, Path	chr8	143740272	T	G	JRK	.	nonsynonymous SNV	X
1, Clin	chr8	143747275	C	G	JRK	.	nonsynonymous SNV	NO, wt/wt
1, Clin	chr8	143745974	G	T	JRK	.	nonsynonymous SNV	NO, wt/wt
<b>1, Clin</b>	<b>chr15</b>	<b>23892819</b>	<b>C</b>	<b>T</b>	<b>MAGEL2</b>	.	<b>nonsynonymous SNV</b>	<b>YES, wt/mt</b>
<b>1, Path</b>	<b>chr15</b>	<b>23892006</b>	<b>C</b>	<b>T</b>	<b>MAGEL2</b>	.	<b>nonsynonymous SNV</b>	<b>X</b>
<b>1, Path</b>	<b>chr15</b>	<b>23892216</b>	<b>G</b>	<b>C</b>	<b>MAGEL2</b>	.	<b>nonsynonymous SNV</b>	<b>X</b>
<b>1, Path</b>	<b>chr15</b>	<b>23892846</b>	<b>G</b>	<b>A</b>	<b>MAGEL2</b>	.	<b>nonsynonymous SNV</b>	<b>X</b>
<b>1, Path</b>	<b>chr15</b>	<b>23889169</b>	<b>C</b>	<b>T</b>	<b>MAGEL2</b>	.	<b>nonsynonymous SNV</b>	<b>X</b>
<b>1, Clin</b>	<b>chr15</b>	<b>23889634</b>	<b>C</b>	<b>A</b>	<b>MAGEL2</b>	.	<b>nonsynonymous SNV</b>	<b>YES, wt/mt</b>
<b>1, Path</b>	<b>chr15</b>	<b>23891698</b>	<b>C</b>	<b>G</b>	<b>MAGEL2</b>	.	<b>nonsynonymous SNV</b>	<b>X</b>
2, Clin	chr1	17721517	G	A	PADI6	.	nonsynonymous SNV	NO, both wt/wt
<b>1, Clin</b>	<b>chr19</b>	<b>1527955</b>	<b>C</b>	<b>T</b>	<b>PLK5</b>	.	<b>nonsynonymous SNV</b>	<b>YES, wt/mt</b>
<b>1, Path</b>	<b>chr19</b>	<b>1528059</b>	<b>G</b>	<b>A</b>	<b>PLK5</b>	.	<b>nonsynonymous SNV</b>	<b>X</b>
<b>1, Clin</b>	<b>chr19</b>	<b>1528003</b>	<b>G</b>	<b>C</b>	<b>PLK5</b>	.	<b>nonsynonymous SNV</b>	<b>YES, wt/mt</b>
1, Path	chr8	145741992	G	A	RECQL4	.	nonsynonymous SNV	X
<b>1, Clin</b>	<b>chr8</b>	<b>145741895</b>	<b>C</b>	<b>T</b>	<b>RECQL4</b>	.	<b>nonsynonymous SNV</b>	<b>YES, mt/mt</b>
<b>1, Clin</b>	<b>chr8</b>	<b>145739330</b>	<b>C</b>	<b>T</b>	<b>RECQL4</b>	.	<b>nonsynonymous SNV</b>	<b>YES, wt/mt</b>
<b>1, Clin</b>	<b>chr8</b>	<b>145740372</b>	<b>C</b>	<b>G</b>	<b>RECQL4</b>	.	<b>nonsynonymous SNV</b>	<b>YES, wt/mt</b>
1, Path	chr8	145738448	A	G	RECQL4	.	nonsynonymous SNV	X
1, Clin	chr8	145739087	T	A	RECQL4	.	nonsynonymous SNV	X
1, Path	chr8	145737667	C	G	RECQL4	.	nonsynonymous SNV	X
1, Path	chr1	182442891	C	A	RGSL1	.	stopgain SNV	X
1, Clin	chr1	182496832	G	C	RGSL1	.	nonsynonymous SNV	YES, wt/mt
1, Path	chr1	182442899	G	C	RGSL1	.	nonsynonymous SNV	X

1, Clin	chr1	182441573	T	A	RGSL1	.	nonsynonymous SNV	NO, wt/wt
1, Clin	chr1	182501804	G	C	RGSL1	.	nonsynonymous SNV	YES, wt/mt
1, Clin	chr1	182443527	G	T	RGSL1	.	nonsynonymous SNV	YES, wt/mt
1, Clin	chr1	182522662	G	C	RGSL1	.	nonsynonymous SNV	YES, wt/mt
1, Path	chr17	19319333	A	G	RNF112	.	nonsynonymous SNV	X
<b>1, Clin</b>	<b>chr17</b>	<b>19318158</b>	<b>C</b>	<b>A</b>	<b>RNF112</b>	.	<b>nonsynonymous SNV</b>	<b>YES, wt/mt</b>
1, Path	chr17	19319390	G	A	RNF112	.	nonsynonymous SNV	X
1, Path	chr17	19314726	C	G	RNF112	.	nonsynonymous SNV	X
1, Path	chr19	53958459	A	G	ZNF761	.	nonsynonymous SNV	X
1, Path	chr19	53959581	A	G	ZNF761	.	nonsynonymous SNV	X
1, Path	chr19	53959131	G	A	ZNF761	.	nonsynonymous SNV	X
<b>1, Clin</b>	<b>chr19</b>	<b>53959466</b>	<b>C</b>	<b>T</b>	<b>ZNF761</b>	.	<b>nonsynonymous SNV</b>	<b>YES, wt/mt</b>
<b>1, Clin</b>	<b>chr19</b>	<b>53958585</b>	<b>G</b>	<b>A</b>	<b>ZNF761</b>	.	<b>nonsynonymous SNV</b>	<b>YES, wt/mt</b>
1, Clin	chr7	6661490	C	T	ZNF853	.	stopgain SNV	NO, wt/wt
1, Path	chr7	6660980	C	G	ZNF853	.	nonsynonymous SNV	X
1, Clin	chr7	6662446	C	A	ZNF853	.	nonsynonymous SNV	NO, wt/wt
1, Clin	chr7	6661799	C	G	ZNF853	.	nonsynonymous SNV	YES, wt/mt
1, Clin	chr9	115759981	G	A	ZNF883	.	nonsynonymous SNV	NO, wt/wt
1, Path	chr9	115760244	C	T	ZNF883	.	nonsynonymous SNV	X

**Table 13: Sanger sequencing results based on gene-based approach analysis.**

An “X” represents insufficient sample available for sanger sequencing validation. Genes in bold are those that confirmed among all variants tested in that gene.

Of the variant positive samples 36 were available for Sanger confirmation and the majority of these novel variants (27) were confirmed (Table 14). Below are two of the Sanger sequencing confirmed results of variants in *CASP8AP2* and *RGSL1*.

```

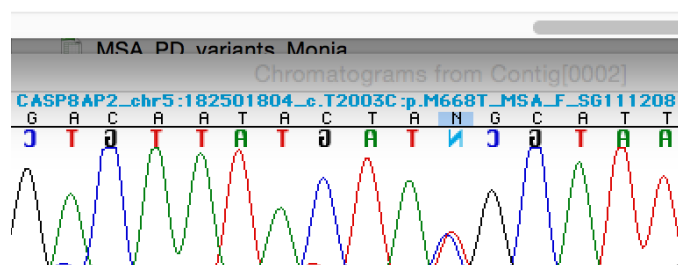
AAATAGTTTGTAGTTAGGTCTGTTGACAACTACTATGCATTGTGAGAGGCCCAT
AAATAGTTTGTAGTTAGGTCTGTTGACAACTACTATGCATTGTGAGAGGCCCAT
AAATAGTTTGTAGTTAGGTCTGTTGACAACTACTATGCATTGTGAGAGGCCCAT
AAATAGTTTGTAGTTAGGTCTGTTGACAACTACTATGCATTGTGAGAGGCCCAT
AAATAGTTTGTAGTTAGGTCTGTTGACAACTACTATGCATTGTGAGAGGCCCAT

```

```

|470|480|490|500|510|
AAATAGTTTGTAGTTAGGTCTGTTGACAACTACTATGCATTGTGAGAGGCCCAT
+
VARIANT_3

```



**Figure 38: Sanger sequencing results in MSA sample heterozygous for *CASP8AP2*, p.M668T**

MSA sample “MSA\_F\_SG111208” was sanger sequenced in several wells to ensure good quality sequence was obtained.

```

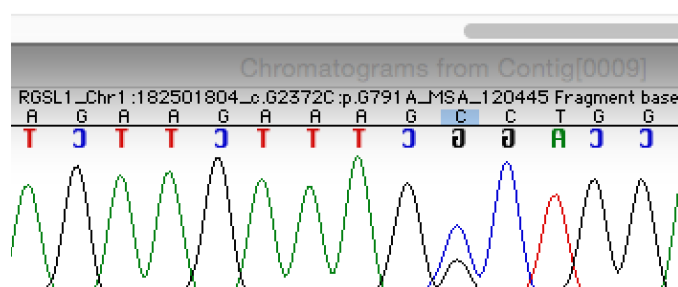
TGTATACCTCCTCTTTCTTTCCCGAGAGAAAGCTGGATGAGATGATC
TGTATACCTCCTCTTTCTTTCCCGAGAGAAAGCTGGATGAGATGATC
TGTATACCTCCTCTTTCTTTCCCGAGAGAAAGCTGGATGAGATGATC
TGTATACCTCCTCTTTCTTTCCCGAGAGAAAGCTGGATGAGATGATC
TGTATACCTCCTCTTTCTTTCCCGAGAGAAAGCTGGATGAGATGATC

```

```

|420|430|440|450|460|
TGTATACCTCCTCTTTCTTTCCCGAGAGAAAGCTGGATGAGATGATC
•
VAR

```



**Figure 39: Sanger sequencing results in MSA sample heterozygous for *RGS1*, p.G791A**

MSA sample “MSA\_120445” was sanger sequenced in several wells to ensure good quality sequence was obtained.

Upon determining which variants were validated, we assessed the prediction filters and excluded 11 variants from our candidate list that were predicted as benign and/or polymorphisms (Table 14). We did not, however, exclude any genes, as there were several other variants within these genes that we were unable to follow-up due to insufficient DNA. We believe that these genes represent strong candidates for containing disease-associated variability and these are thus included in the hypothesis generating result set.

	Number of samples , Clin or Path	#CHROM	POSITION	REF	ALT	GENE NAME	ID	Sanger Confirmation, Genotype(s)	Prediction	
2	1, Clin	chr22	18121557	C	T	BCL2L13	-	nonsynonymous SNV	YES, wt/mt	disease causing/very damaging
3	1, Clin	chr22	18121587	G	T	BCL2L13	-	nonsynonymous SNV	YES, wt/mt	polymorphism;benign
4	1, Clin	chr6	90573431	T	C	CASP8AP2	-	nonsynonymous SNV	YES, wt/mt	unknown
5	1, Clin	chr6	90577775	A	G	CASP8AP2	-	nonsynonymous SNV	YES, wt/mt	unknown
6	1, Clin	chr6	90572365	T	G	CASP8AP2	-	nonsynonymous SNV	YES, wt/mt	unknown
7	1, Clin	chr11	11373751	G	A	CSNK2A3	-	stopgain SNV	YES, wt/mt	unknown
8	1, Clin	chr11	11373910	C	A	CSNK2A3	-	nonsynonymous SNV	YES, wt/mt	unknown
9	1, Clin	chr3	53322243	G	A	DCP1A	-	nonsynonymous SNV	YES, wt/mt	unknown
10	1, Clin	chr3	53376290	G	A	DCP1A	-	nonsynonymous SNV	YES, wt/mt	unknown
11	1, Clin	chr7	1786631	T	G	ELFN1	-	nonsynonymous SNV	YES, wt/mt	disease causing/very damaging
12	1, Clin	chr7	1785179	G	A	ELFN1	-	nonsynonymous SNV	YES, wt/mt	polymorphism;benign
13	2, Clin	chr7	150434680	G	A	GIMAP1-GIMAP5	-	nonsynonymous SNV	YES, both wt/mt	polymorphism;splicing affected
14	1, Clin	chr15	23892819	C	T	MAGEL2	-	nonsynonymous SNV	YES, wt/mt	unknown
15	1, Clin	chr15	23889634	C	A	MAGEL2	-	nonsynonymous SNV	YES, wt/mt	polymorphism;benign
16	1, Clin	chr19	1527955	C	T	PLK5	-	nonsynonymous SNV	YES, wt/mt	polymorphism;benign
17	1, Clin	chr19	1528003	G	C	PLK5	-	nonsynonymous SNV	YES, wt/mt	disease causing/very damaging; splicing affected
18	1, Clin	chr8	145741895	C	T	RECQL4	-	nonsynonymous SNV	YES, mt/mt	unknown
19	1, Clin	chr8	145739330	C	T	RECQL4	-	nonsynonymous SNV	YES, wt/mt	unknown
20	1, Clin	chr8	145740372	C	G	RECQL4	-	nonsynonymous SNV	YES, wt/mt	unknown
21	1, Clin	chr1	182496832	G	C	RGSL1	-	nonsynonymous SNV	YES, wt/mt	polymorphism;benign
22	1, Clin	chr1	182501804	G	C	RGSL1	-	nonsynonymous SNV	YES, wt/mt	polymorphism;splicing affected
23	1, Clin	chr1	182443527	G	T	RGSL1	-	nonsynonymous SNV	YES, wt/mt	polymorphism;benign
24	1, Clin	chr1	182522662	G	C	RGSL1	-	nonsynonymous SNV	YES, wt/mt	polymorphism;benign
25	1, Clin	chr17	19318158	C	A	RNF112	-	nonsynonymous SNV	YES, wt/mt	polymorphism;benign
26	1, Clin	chr19	53959466	C	T	ZNF761	-	nonsynonymous SNV	YES, wt/mt	unknown
27	1, Clin	chr19	53958585	G	A	ZNF761	-	nonsynonymous SNV	YES, wt/mt	unknown
28	1, Clin	chr7	6661799	C	G	ZNF853	-	nonsynonymous SNV	YES, wt/mt	polymorphism;benign

**Table 14: All 27 sanger sequencing confirmed variants from WES with Mutation Taster prediction.**

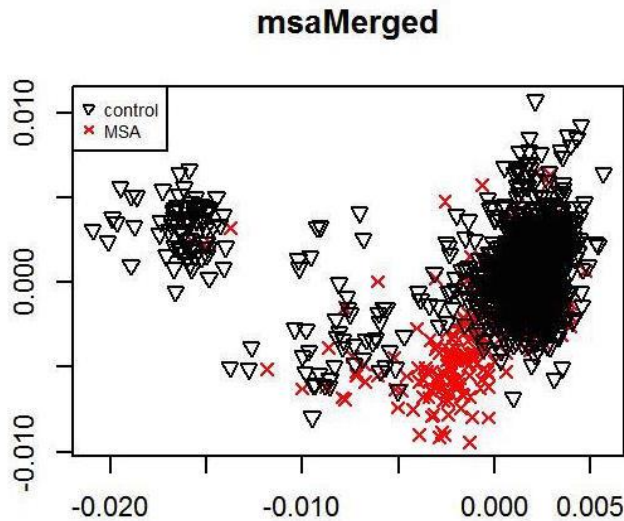
If predictions were unknown, variants remained in our candidate list.

### 3.3.5 RAREMETAL Individual Variant and Gene Burden Analyses

#### 3.3.5.1 Quality Control with GoogleGenome data in GoogleCloud

In our second case/control analysis using GoogleCloud, the first two PCA covariates (C1, C2) were plotted as eigenvectors to assess population stratification among the full cohort (Figure 40). The results demonstrated that all MSA cases cluster uniformly

with study controls, indicating common ancestry and the absence of ethnic outliers. Any MSA cases or controls that deviated from this cluster were removed from further analyses, leaving us with 335 MSA cases and 1085 controls.

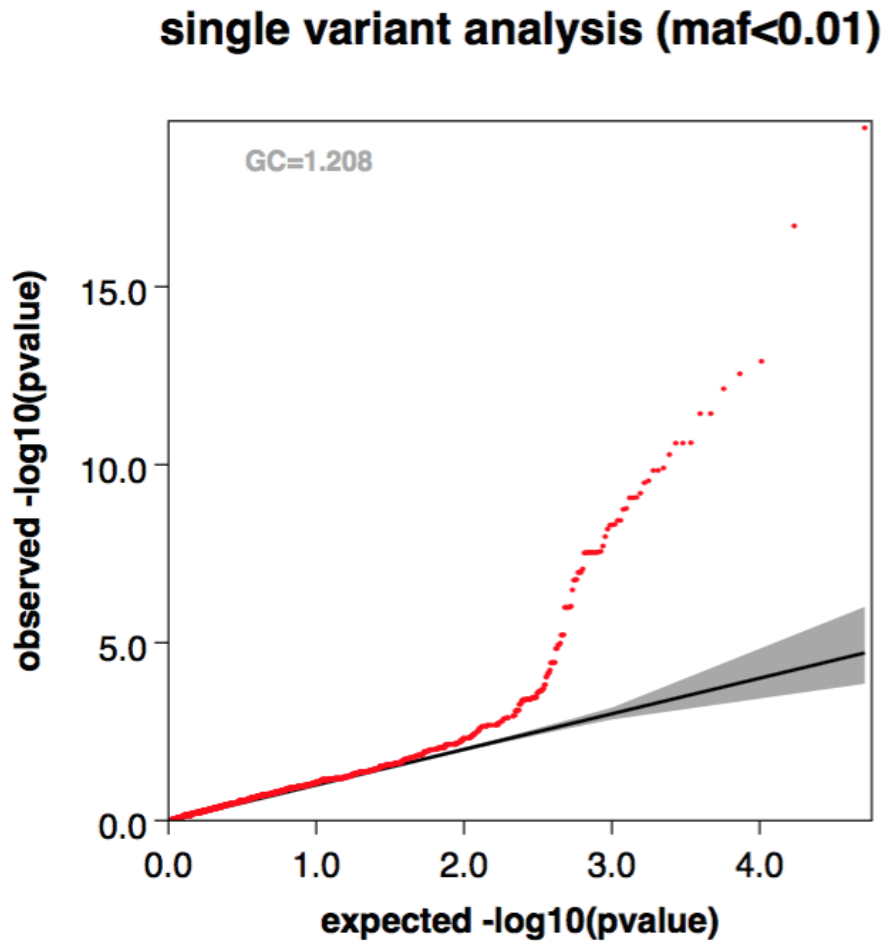


**Figure 40: Population stratification of MSA cases and controls**

We performed multidimensional scaling on the MSA Merged cohort consisting of ~1300 controls and ~400 cases using the first two covariates (C1, C2).

As a second quality control parameter, QQ plots were obtained for each statistical test performed. Results of single variant tests for a  $MAF < 0.01$  are shown below in Figure 41. This was a good example of a QQ plot that deviates from the  $x=y$  axis along the latter half of the graph, indicating little population stratification but the presence of highly significant and rare variants. The genomic inflation factor, depicted by lambda ( $\lambda_{gc}$ ), is used as an additional QC measure to ensure an absence of sample duplications, unknown familial relationships, population stratification and systematic technical bias ([http://rstudio-pubs-static.s3.amazonaws.com/9743\\_8a5f7ba3aa724d4b8270c621fdf6d06e.html](http://rstudio-pubs-static.s3.amazonaws.com/9743_8a5f7ba3aa724d4b8270c621fdf6d06e.html)). With a  $\lambda$

value  $>1$  ( $GC = 1.208$ ,  $GC = 1.164$ , Figure 41, Figure 42, respectively) this suggests slight systematic technical bias, as all of the other possibilities were ruled out using previous QC PCA and IBD analyses and QQ plots.



**Figure 41: QQ plot of single variant results with a MAF<0.01**

$GC = 1.208$  represents the genomic inflation factor,  $\lambda$ .

All burden tests (MB, VT, CMC) results for MAF<0.01 revealed similar QQ plots, with the CMC QQ plot illustrated below (Figure 42). Resonating with the single

variant test results, the  $x=y$  deviation on the right side of the graph indicated the presence of significant variants with minimal population stratification.

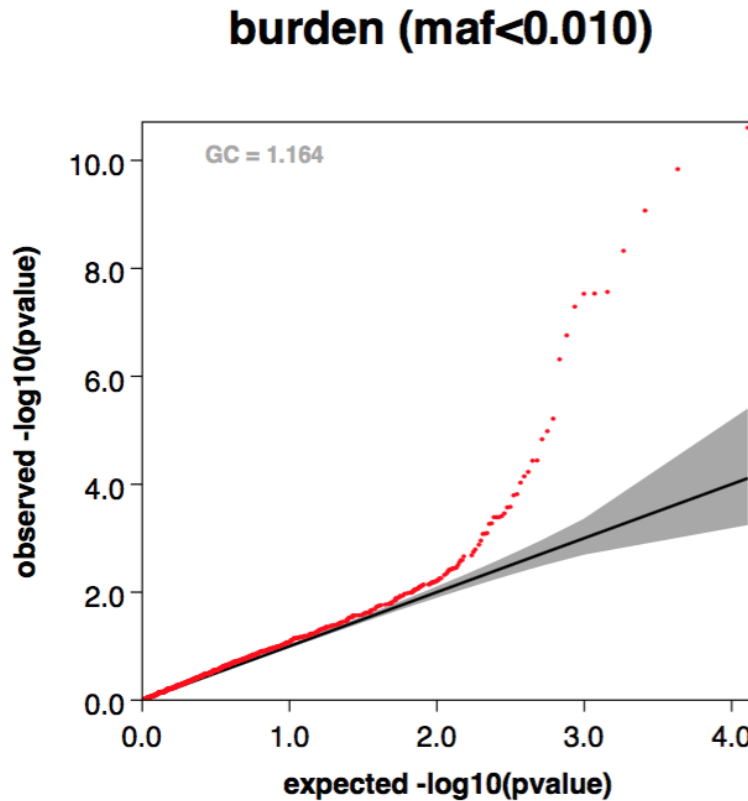


Figure 42: QQ plot of CMC burden test results with a MAF<0.01

GC =1.164 represents the genomic inflation factor,  $\lambda$ .

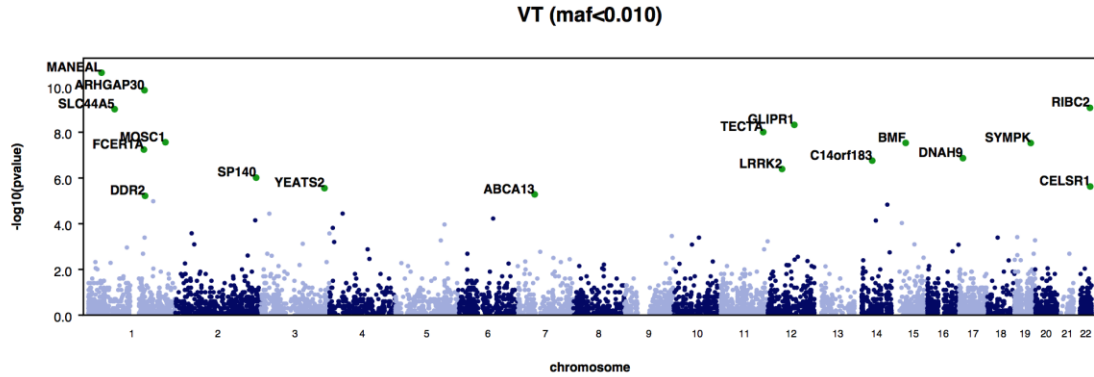
Once we could confirm the results passed quality control filters, we pursued gene burden and individual variant analyses.

### 3.3.5.2 Gene Burden Results

There was substantial overlap among the vast majority of significant genes between burden results from the three different tests (MB, VT, CMC). All burden tests were merged to determine genes with a p-value  $< 1 \times 10^{-6}$  among coding, non-synonymous variants with MAF<0.01 (Table 15). A Manhattan plot of the VT gene



burden test results including coding, non-synonymous variants with a MAF<0.01 is shown below in Figure 43.



**Figure 43: Manhattan plot of VT gene burden test results in coding alleles with a MAF<0.01.**

As a burden test, the VT test assumes that all rare variants within a particular region influence the phenotype in the same direction and magnitude. Genes that illustrate the greatest significance, indicated by  $-\log_{10}(\text{pvalue})$  on the x-axis (i.e. *MANEAL*, *ARHGAP30*, *RIBC2*), adhere at MAF<1%.

GENE	NUM_VAR	AVG_AF	MIN_AF	MAX_AF	EFFECT_SIZE	PVALUE
<i>MANEAL</i>	1	0.00453594	0.00453594	0.00453594	0.779577	2.48E-11
<i>ARHGAP30</i>	1	0.00418702	0.00418702	0.00418702	0.77903	1.45E-10
<i>RIBC2</i>	1	0.0038381	0.0038381	0.0038381	0.0481362	8.50E-10
<i>SLC44A5</i>	2	0.00489169	0.00488486	0.00489853	0.497615	9.76E-10
<i>IGSF21</i>	15	0.00138551	0.000352113	0.0098661	0.327289	2.13E-09
<i>KRT18</i>	11	0.00288117	0.000352113	0.00809859	0.171798	2.34E-09
<i>GLIPR1</i>	1	0.00349406	0.00349406	0.00349406	0.0459683	4.74E-09
<i>TECTA</i>	1	0.00349406	0.00349406	0.00349406	0.778324	9.79E-09
<i>MOSC1</i>	1	0.00314685	0.00314685	0.00314685	0.779029	2.74E-08
<i>BMF</i>	1	0.00314027	0.00314027	0.00314027	0.777387	2.93E-08
<i>SYMPK</i>	1	0.00314465	0.00314465	0.00314465	0.777075	2.97E-08
<i>DNAH9</i>	5	0.00593234	0.00314027	0.0094208	0.252219	5.13E-08
<i>FCER1A</i>	1	0.00314246	0.00314246	0.00314246	0.777935	5.70E-08
<i>C14orf183</i>	1	0.00279525	0.00279525	0.00279525	0.0409978	1.75E-07
<i>ZNF519</i>	25	0.00140845	0.000352113	0.00598592	0.195777	2.16E-07
<i>DRD5</i>	6	0.00187798	0.000352113	0.00739437	0.382788	2.28E-07
<i>SP140</i>	2	0.0047104	0.00279135	0.00662945	0.0285598	3.00E-07
<i>LRRK2</i>	2	0.0038381	0.00348918	0.00418702	0.428455	4.05E-07
<i>OR8D2</i>	12	0.000586854	0.000352113	0.00140845	0.0108746	4.57E-07
<i>DDR2</i>	13	0.000568855	0.000352113	0.00211268	0.457547	4.89E-07
<i>COL3A1</i>	24	0.00098326	0.000352113	0.00633803	1.25E+06	5.14E-07
<i>GANC</i>	20	0.000721831	0.000352113	0.00352113	942612	7.27E-07
<i>DYRK2</i>	6	0.00111519	0.000352113	0.00422535	0.0156217	8.31E-07

**Table 15: The most significant genes with p-values <1X10<sup>-6</sup> among all 3 gene burden tests (MB, VT, CMC).**

Genes highlighted in red text were investigated for further analysis in the next section.

### 3.3.5.2.1 In-depth gene analysis

As discussed above, the primary aim of this work was to generate a list of candidate genes and variants in MSA; and the list above in Table 15 qualifies; however, it is worth discussing some of these candidates because the previous involvement in neurological disease, and in particular in synucleinopathies, is striking.

Therefore for several genes, we performed a more focused analysis of protein altering coding variants based on their functional relevance to PD. Some genes harboring significant burdens carry variants known to cause monogenic forms of familial PD, including *LRRK2* and *PARK2*. Other genes have been reported to manifest an association with PD, such as *EIF4G1* and *GIGYF2*; these, however, are tentative, as independent replication is lacking or controversial. Other genes closely analyzed all demonstrate substantial neuronal expression, particularly in the cortex and cerebellum, making these good candidates for further exploration of non-synonymous variants and their potential role as risk or protective factors.

#### 3.3.5.2.1.1 Genes associated with monogenic forms of familial PD

##### *LRRK2*

CHROM	POS	ID	REF	ALT	Effect	MAF	Prediction (polyphen, SIFT)	Allele Freq: cases	Allele Freq: controls
12	40629436	rs33995463	T	C	Nonsynonymous	0.00129	unknown	T:738, C:4	T:2226, C:6
12	40657700	rs7308720	C	G	Nonsynonymous	0.08607	unknown	C:692, G:50	C:2082, G:150
12	40671989	rs10878307	A	G	Nonsynonymous	0.05922	unknown	A:700, G:42	A:2047, G:185
12	40702911	rs7133914	G	A	Nonsynonymous	0.08412	benign;tolerated	G:690, A:52	G:2082, A:150
12	40707778	rs35507033	G	A	Nonsynonymous	0.00379	benign;tolerated	G:740, A:0	G:2213, A:17
12	40707861	rs33958906	C	T	Nonsynonymous	0.03011	possibly damaging; deleterious	C:717, T:25	C:2151, T:81
12	40713899	rs35303786	T	C	Nonsynonymous	0.00916	benign;tolerated	T:735, C:7	T:2204, C:28
12	40713901	rs11564148	T	A	Nonsynonymous	0.2983	benign;tolerated	T:534, A:208	T:1555, A:677
12	40740686	rs33995883	A	G	Nonsynonymous	0.01764	probably damaging; deleterious	A:731, G:11	A:2197, G:35
12	40757328	.	G	C	Nonsynonymous	.	unknown	G:730, C:12	G:2232, C:0

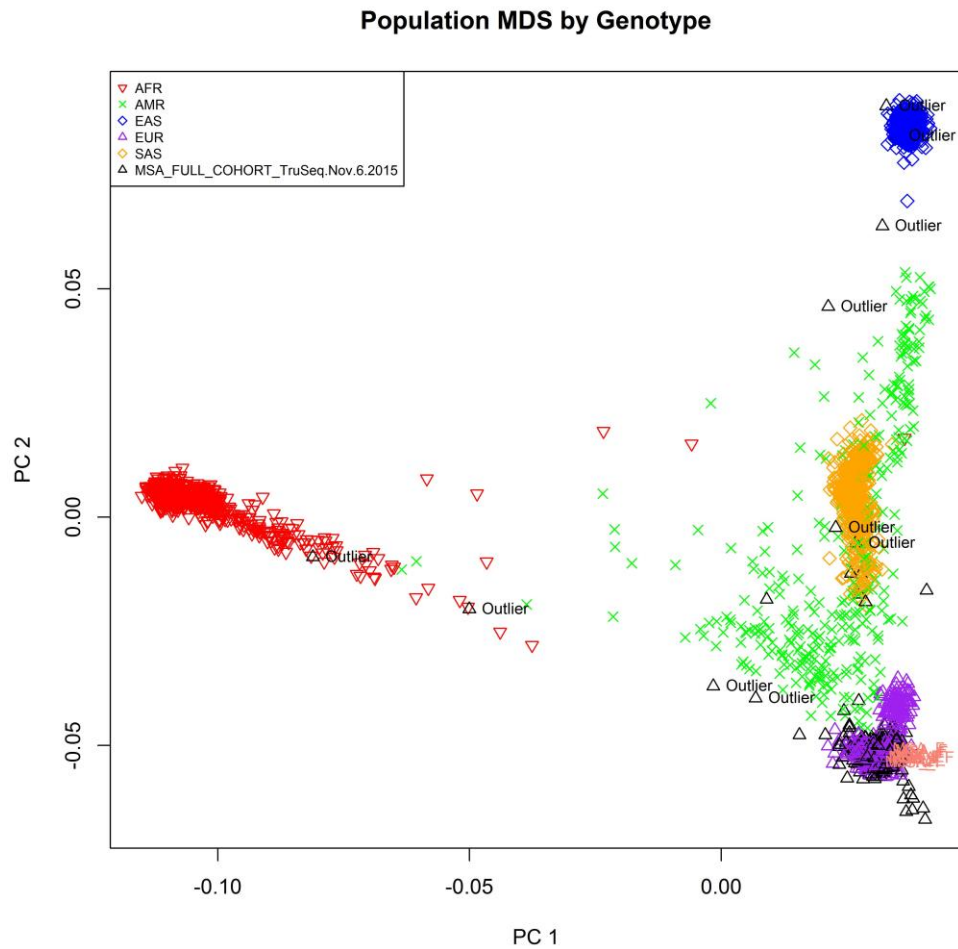
**Table 16: Non-synonymous *LRRK2* variants identified by gene burden analyses.**

Controls: n = 1085. Cases: n =335. Text highlighted in blue reflects variants with allele counts absent in cases but present in control only. Text highlighted in red reflects variants with counts absent in controls but present in cases only.

Given the approximate 1:3.25 ratio of cases to controls, respectively, we would anticipate the number of both major and minor alleles in controls to be roughly 3.25X that of cases. In Table 16, there are two alleles that largely deviate from this expected ratio. Highlighted in blue, the minor allele for rs3550733 is present in 17 controls and would be expected to be in approximately 4 cases. While it is absent in all cases, this suggests the possibility that this minor allele imparts a protective effect against the development of PD.

Highlighted in red is variant p.G2385R, which is present in 10 MSA cases, two of which are homozygous. While this variant has been identified before, it is exclusive to Asian populations.<sup>173,174,181</sup> As our cases are all of Caucasian ancestry, as determined by MDS uniform clustering of covariates C1 and C2, this is a very unique finding (Figure 40).

To verify that these 10 individuals were indeed of Caucasian ancestry, we plotted them with a unique coral color to identify their location on the MDS plot (Figure 44).



**Figure 44: MDS plot of MSA samples with 10 individuals carrying LRRK2 p.G2385R in coral.**

Individuals with LRRK2 p.G2385R mutation are represented by the coral colored triangles on the lower right side of the plot. All 10 individuals tightly cluster with those of European ancestry.

**Superpopulations key:**

- **AFR**, African
- **AMR**, Mixed American
- **EAS**, East Asian
- **EUR**, European
- **SAS**, South Asian

From the MDS plot it is evident that these 10 individuals cluster tightly with those of European ancestry, suggesting a novel finding: this variant is present among Caucasian populations and not exclusive to Asian populations.

To confirm that this variant was indeed real, we followed-up with Sanger sequencing on those 10 individuals. However, this variant could not be validated and therefore must be attributed to miscalling. On review of the underlying short read sequence there is an imbalance of the variant allele being at a lower coverage than the wild type allele, and thus we believe that this is an artifact of the variant calling process.

### PARK2

#CHROM	POS	ID	REF	ALT	Effect	MAF	Prediction (polyphen, SIFT)	Allele Freq: cases	Allele Freq: controls
6	161771219	rs149953814	G	A	Nonsynonymous	0.001495	unknown	G:739, A:3	G:2226, A:6
6	161781225	rs1801334	C	T	Nonsynonymous	0.0256	unknown	C:707, T:33	C:2155, T:77
6	161807855	rs1801582	C	G	Nonsynonymous	0.1646	unknown	C:595, G:147	C:1885, G:347
6	162206852	rs34424986	G	A	Nonsynonymous	0.002059	unknown	G:739, A:3	G:2223, A:9
6	162622197	rs1801474	C	T	Nonsynonymous	0.06758	unknown	C:729, T:13	C:2192, T:40
6	162683724	rs55774500	G	T	Nonsynonymous	0.00472	unknown	G:741, T:1	G:2227, T:5

**Table 17: Non-synonymous *PARK2* variants identified by gene burden analyses.**

Controls: n = 1085. Cases: n =335

### 3.3.5.2.1.2 Genes with tentative PD associations

### EIF4GI

#CHROM	POS	ID	REF	ALT	Effect	MAF	Prediction (polyphen, SIFT)	Allele Freq: cases	Allele Freq: controls
3	184039766	rs111659103	AAGGAGAAGC	A	Codon Loss/Inframe Deletion	0.01729	unknown	AAGGAGAAGC:727, A:15	AAGGAGAAGC:2172, A:60
3	184040606	rs190378563	C	T	Normal_splice_site	0.001688	unknown	C:736, T:6	C:2225, T:7
3	184045397	rs35629949	C	G	Nonsynonymous	0.003153	unknown	C:737, G:5	C:2221, G:11
3	184045410	rs2230570	T	C	Nonsynonymous	0.02257	unknown	T:726, C:16	T:2172, C:60

**Table 18: Non-synonymous *EIF4GI* variants identified by gene burden analyses.**

Controls: n = 1085. Cases: n =335

### GIGYF2

#CHROM	POS	ID	REF	ALT	Effect	MAF	Prediction (polyphen, SIFT)	Allele Freq: cases	Allele Freq: controls
2	233659553	rs2289912	C	A	Nonsynonymous	0.05238	unknown	C:732, A:10	C:2192, A:40
2	233712109	rs72554081	A	G	Nonsynonymous	0.001573	unknown	A:739, G:3	A:2228, G:4

**Table 19: Non-synonymous *GIGYF2* variants identified by gene burden analyses.**

Controls: n = 1085. Cases: n =335

### 3.3.5.2.1.3 Functionally interesting genes for further exploration

As *VPSI3C* has recently been identified to be associated with PD, we decided to investigate non-synonymous variants in sister gene, *VPSI3D*.<sup>342,343</sup> However, none of the variants revealed significant differences in allele frequencies between cases and controls (Table 20).

#### *VPSI3D*

#CHROM	POS	ID	REF	ALT	Effect	MAF	Prediction (polyphen, SIFT)	Allele Freq: cases	Allele Freq: controls
1	12313827	rs61774897	A	G	Nonsynonymous	0.002463	unknown	A:738, G:4	A:2221, G:11
1	12339619	rs4845898	A	T	Nonsynonymous	0.2162	unknown	A:564, T:178	A:1702, T:530
1	12342990	rs41279452	G	A	Nonsynonymous	0.009227	unknown	G:737, A:5	G:2201, A:31
1	12401857	rs143194636	C	T	Nonsynonymous	0.004231	unknown	C:741, T:1	C:2226, T:6

**Table 20: Non-synonymous *VPSI3D* variants identified by gene burden analyses.**

Controls: n = 1085. Cases: n =335

For the next two genes, *SLC44A5* and *GLIPR1*, there was a single non-synonymous variant in each gene that is only present in several pathologically confirmed cases (16, 11, respectively) but absent in controls. We pursued Sanger sequencing with both of these variants (highlighted in red below); however, neither were successfully validated.

#### *SLC44A5*

#CHROM	POS	ID	REF	ALT	Effect	MAF	Prediction (polyphen, SIFT)	Allele Freq: cases	Allele Freq: controls
1	75684184	rs143004355	C	T	Nonsynonymous	0.003103	unknown	C:734, T:6	C:2224, T:8
1	75684185	rs200948281	G	C	Nonsynonymous	0.008846	unknown	G:716, C:16	G:2232, C:0
1	75688101	rs138027438	T	G	Nonsynonymous	0.007659	unknown	T:735, G:7	T:2203, G:29
1	75707678	rs141693552	C	T	Normal splice site	0.01771	unknown	C:732, T:10	C:2191, T:41
1	75862263	rs79353172	C	T	Normal splice site	0.004512	unknown	C:738, T:4	C:2211, T:21

**Table 21: Non-synonymous *SLC44A5* variants identified by gene burden analyses.**

Controls: n = 1085. Cases: n =335. Text highlighted in red reflects variants with counts absent in controls but present in cases only.

*GLIPR1*

#CHROM	POS	ID	REF	ALT	Effect	MAF	Prediction (polyphen, SIFT)	Allele Freq: cases	Allele Freq: controls
12	75785047	rs146467128	G	A	Nonsynonymous	0.002586	unknown	G:737, A:5	G:2224, A:8
12	75816814	rs144850646	G	GACA	CodonGain/Inframe Insertion	0.3618	unknown	G:451, GACA:291	G:1368, GACA:864
12	75874811	rs76259505	A	C	Nonsynonymous	0.003885	unknown	A:727, C:11	A:2232, C:0

**Table 22: Non-synonymous *GLIPR1* variants identified by gene burden analyses.**

Controls: n = 1085. Cases: n =335. Text highlighted in red reflects variants with counts absent in controls but present in cases only.

While the non-synonymous variant highlighted below in *CASP8AP2* is present in controls, the doubled allele frequency in cases (notably from separate geographic cohorts (French, UK, USA)), despite 1:3.25 size ratio of total case to control cohort, suggests a possible MSA risk allele. Furthermore, the rarity of this allele (MAF=.0012 on the Exac database) provides further support as a plausible MSA risk variant.

*CASP8AP2*

#CHROM	POS	ID	REF	ALT	Effect	MAF	Prediction (polyphen, SIFT)	Allele Freq: cases	Allele Freq: controls
6	90577706	.	TG	T	Frameshift	.	unknown	TG:655, T:87	TG:1942, T:290
6	90577711	.	TCTTTGCCCGACATGGA	T	Frameshift	.	unknown	TCTTTGCCCGACATGGA:655, T:87	TCTTTGCCCGACATGGA:1942, T:290
6	90577876	rs150022229	G	T	Nonsynonymous*	0.001198	unknown	G:736, T:6	G:2229, T:3

**Table 23: Non-synonymous *CASP8AP2* variants identified by gene burden analyses.**

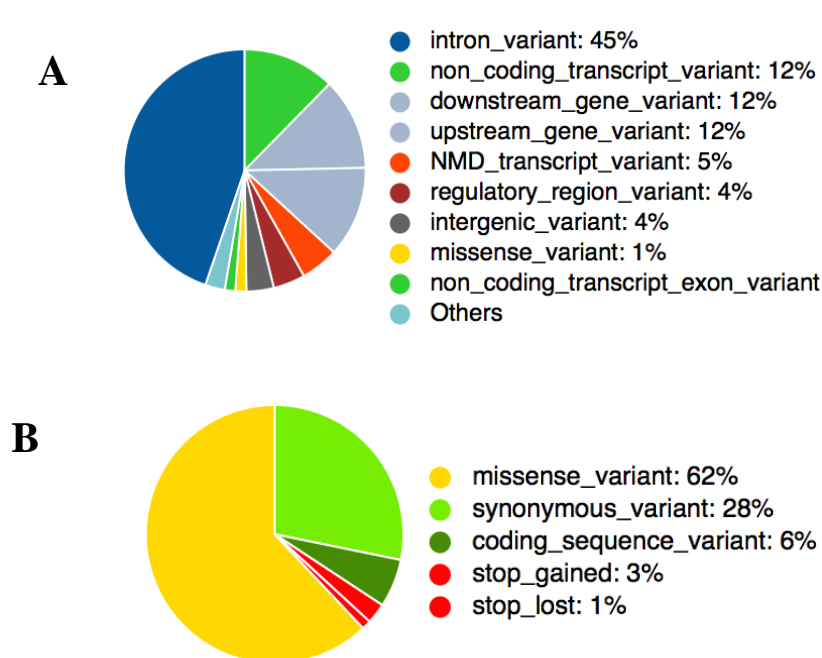
Controls: n = 1085. Cases: n =335. Text highlighted in red reflects variants with counts greater in cases than controls.

Using Sanger sequencing, we were able to confirm this allele in all 5 of the 5 individuals tested. Adequate DNA sample was unavailable for the 6<sup>th</sup> individual.

*3.3.5.3 Single Variant Results**3.3.5.3.1 Overview of Results*

As RAREMETAL generates large quantities of data using several different tests and models, it was helpful to visually categorize this information. All single variant results were run through the “Variant Effect Predictor” tool on Ensembl (<http://www.ensembl.org>).

ensembl.org). The consequences of all data and coding data, respectively, are illustrated below.

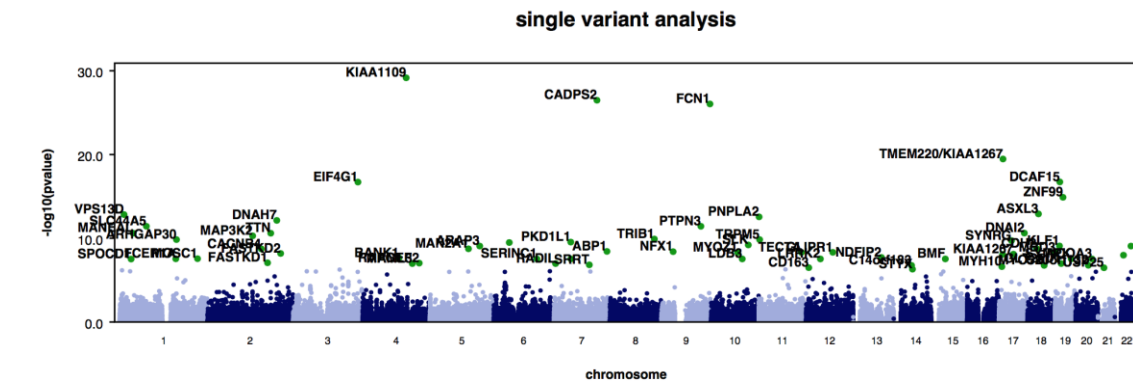


**Figure 45: Ensembl Variant Effect Predictor Tool results of individual variant consequences.**

- A. Consequences of all individual variants
- B. Consequence of coding variants only

Among the coding variants, the largest proportion consisted of non-synonymous missense mutations. A Manhattan plot of the most significant single variant hits is shown below in Figure 46. This includes both coding and non-coding variants with no MAF cut-off.

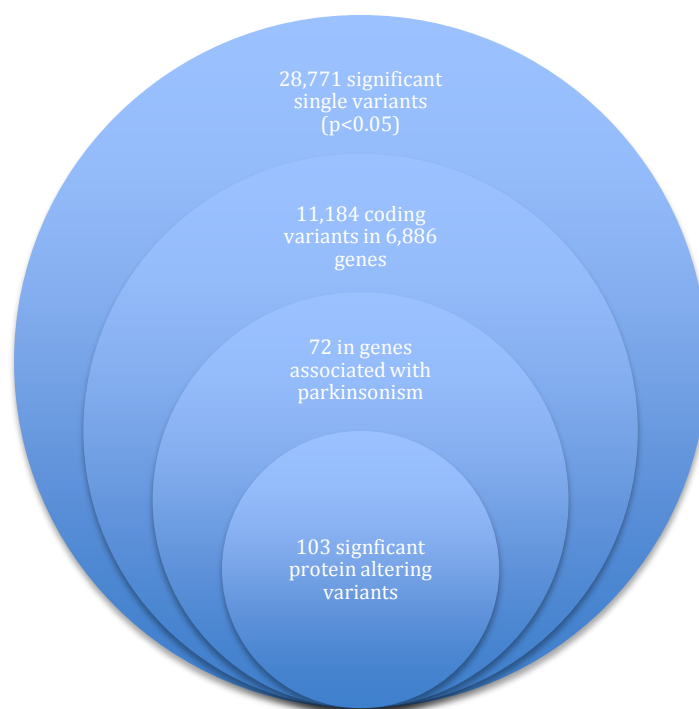




**Figure 46: Manhattan plot of top hits from single variant analyses in MSA exome cohort.**

This includes all variants (coding, non-coding) with no MAF cut-off. Unlike a burden analysis, a single variant analysis does not assume that variants are causal or that other variants within the same gene will exhibit the same direction and magnitude of association. Notably, some of the single variants with very significant p values, as demonstrated by  $-\log_{10}(\text{pvalue})$  on the x-axis, are located in genes that have been suggested to harbor associations with non-monogenic forms of PD (i.e. *VPS13D* and *EIF4G1*, respectively).

Using a p-value  $< 0.05$  as a cut-off, a total of 28,771 significant variants were generated. As we were interested in coding variants, we filtered the list down to 11,184 coding variants in 6,886 genes. All genes were run through functional annotation clustering in DAVID (<https://david.ncifcrf.gov/>), and results for genes associated with parkinsonism were extracted, yielding 72 genes, some of which harbor variants responsible for monogenic forms of PD (*LRRK2*, *PARK2*). As we are interested in variants that alter protein coding, we focused on non-synonymous variants. Among these 72 genes there were 103 single variants (non-synonymous, stopgain, stoploss, splice site, codon loss/inframe deletion) with p-values  $< 0.05$  (Figure 47, Table 24). Among all non-synonymous single variants associated with PD, *LRRK2* had the most significant p-value:  $2.47 \times 10^{-6}$ .



**Figure 47: Filtering results of single variant burden analyses for MSA exome cohort**

Each circle reflects subsequent stages of filtering after applying the appropriate exclusion criteria.

CHR	POS	REF	ALT	POOLED ALT_AF	EFFECT SIZE	EFFECT_SI ZE SD	PVALUE	Effect	Gene
12	40757328	G	C	0.00316901	0.542968	0.115265	0.00000247	NS	<i>LRRK2</i>
4	9783915	C	T	0.00739437	0.35579	0.0910518	0.0000932	NS	<i>DRD5</i>
8	26721808	A	G	0.00140845	0.788407	0.206159	0.000131162	NS	<i>ADRA1A</i>
1	161180187	G	A	0.0028169	0.542024	0.146842	0.000223193	NSS	<i>NDUFS2</i>
6	31918464	A	G	0.00985915	0.250116	0.0735243	0.00066939	NS	<i>CFB</i>
2	207012392	T	C	0.0056338	0.351534	0.103818	0.000709007	NSS	<i>NDUFS1</i>
3	49394834	G	A	0.319366	-0.0552046	0.0165917	0.000877113	NS	<i>GPX1</i>
11	83673931	G	A	0.00105634	0.785639	0.238374	0.000981339	NS	<i>DLG2</i>
15	89873481	A	C	0.00105634	0.778453	0.242174	0.00130702	NS	<i>POLG</i>
10	135351264	G	A	0.00176056	0.583736	0.185317	0.00163306	NS	<i>CYP2E1</i>
3	133494354	C	T	0.0926056	-0.0768987	0.0267417	0.00403246	NS	<i>TF</i>
12	117725949	C	T	0.000704225	0.786227	0.291613	0.00701491	NS	<i>NOS1</i>

3	49848470	C	T	0.000704722	0.790035	0.293143	0.00703775	NS	<i>UBA7</i>
2	25384092	T	C	0.00140845	0.549644	0.204416	0.00716982	NS	<i>POMC</i>
2	240960719	C	T	0.000704225	0.783499	0.293069	0.00750803	NS	<i>NDUFA10</i>
4	9783969	T	C	0.000704225	0.783499	0.293069	0.00750803	NS	<i>DRD5</i>
4	68424646	G	A	0.000704722	0.784214	0.293367	0.00751432	NS	<i>STAP1</i>
2	219678877	C	T	0.0264085	-0.128627	0.0483048	0.00774912	NS	<i>CYP27A1</i>
8	31498066	G	A	0.000704722	0.780301	0.295186	0.0082073	NS	<i>NRG1</i>
16	72108188	C	G	0.000704225	0.778972	0.295185	0.00831693	NS	<i>HPR</i>
16	72108192	T	G	0.000704225	0.778972	0.295185	0.00831693	NS	<i>HPR</i>
19	39384491	G	C	0.000704225	0.778972	0.295185	0.00831693	NS	<i>SIRT2</i>
3	48638440	C	T	0.000704225	0.776659	0.296257	0.00875265	NS	<i>UQCRC1</i>
3	49850571	C	T	0.000704225	0.776659	0.296257	0.00875265	NS	<i>UBA7</i>
6	75953484	G	C	0.000704225	0.776659	0.296257	0.00875265	NS	<i>COX7A2</i>
3	49846412	A	C	0.00140845	0.534892	0.204591	0.00893735	NS	<i>UBA7</i>
7	100488638	G	T	0.00211715	0.446006	0.172687	0.00980183	NS	<i>ACHE</i>
4	9784658	C	A	0.00140845	0.532013	0.208248	0.0106277	SG	<i>DRD5</i>
12	112221070	G	C	0.00140845	0.525689	0.206262	0.0108144	NS	<i>ALDH2</i>
10	88719758	T	C	0.00140845	0.525839	0.206568	0.0109089	NS	<i>SNCG</i>
2	233655834	T	A	0.00140845	0.524278	0.206533	0.011134	NS	<i>GIGYF2</i>
6	170871037	GCA A	G	0.25689	0.0432992	0.0172904	0.0122716	CL	<i>TBP</i>
12	9243017	G	A	0.0257223	0.119384	0.047943	0.0127699	NS	<i>A2M</i>
10	88722398	A	T	0.240493	0.0448042	0.0181079	0.0133502	NS	<i>SNCG</i>
1	196712596	A	T	0.0179577	-0.141341	0.058044	0.0148893	NS	<i>CFH</i>
8	27358505	A	G	0.102817	-0.0619026	0.0255627	0.0154525	NS	<i>EPHX2</i>
14	55310492	G	A	0.197248	-0.0457533	0.0195964	0.0195547	NSS	<i>GCH1</i>
12	40707778	G	A	0.00599436	-0.222286	0.10102	0.0277772	NS	<i>LRRK2</i>
6	31778077	T	G	0.0302817	0.0987488	0.0455407	0.0301312	NS	<i>HSPAIL</i>
16	2134221	C	T	0.00634249	0.189988	0.0886652	0.0321322	NSS	<i>TSC2</i>
19	6702598	A	G	0.248239	-0.0378569	0.0177075	0.0325243	NSS	<i>C3</i>
3	49847804	C	T	0.00457746	0.247008	0.115749	0.0328424	NS	<i>UBA7</i>
6	161807855	C	G	0.16338	0.0433567	0.0208017	0.0371342	NS	<i>PARK2</i>
14	64700045	T	C	0.031338	0.0934462	0.0452254	0.0388065	NSS	<i>ESR2</i>
1	218610682	C	A	0.00176429	0.379943	0.184587	0.0395582	NSS	<i>TGFB2</i>

15	101565029	G	A	0.00528169	-0.219614	0.107036	0.0401912	NS	<i>LRRK1</i>
21	35281421	T	C	0.105986	-0.0511246	0.025055	0.0413014	NS	<i>ATP5O</i>
7	24324879	T	C	0.0376761	-0.0822546	0.040744	0.0435068	NS	<i>NPY</i>
6	162206909	G	A	0.00176056	0.379482	0.188789	0.0444215	NS	<i>PARK2</i>
5	149456964	T	A	0.00105634	0.47195	0.234982	0.0445949	NS	<i>CSF1R</i>
17	31618865	A	C	0.000352361	0.802158	0.403377	0.0467449	NS	<i>ACCN1</i>
1	226026406	A	G	0.186268	-0.039587	0.0199109	0.0467889	NS	<i>EPHX1</i>
1	16134056	A	T	0.000352113	0.799286	0.403375	0.047536	NS	<i>UQCRL</i>
1	42048868	G	A	0.000352113	0.799286	0.403375	0.047536	NS	<i>HIVEP3</i>
1	54610305	A	C	0.000352113	0.799286	0.403375	0.047536	NS	<i>CDCP2</i>
1	196709872	C	A	0.000352113	0.799286	0.403375	0.047536	NS	<i>CFH</i>
2	25384116	G	A	0.000352113	0.799286	0.403375	0.047536	NS	<i>POMC</i>
2	25384219	C	G	0.000352113	0.799286	0.403375	0.047536	NS	<i>POMC</i>
2	207011682	C	T	0.000352113	0.799286	0.403375	0.047536	NS	<i>NDUFS1</i>
4	6304142	G	A	0.000352113	0.799286	0.403375	0.047536	NS	<i>WFS1</i>
4	100057768	T	G	0.000352113	0.799286	0.403375	0.047536	NS	<i>ADH4</i>
4	100201384	C	T	0.000352113	0.799286	0.403375	0.047536	NS	<i>ADH1A</i>
4	100205629	G	A	0.000352113	0.799286	0.403375	0.047536	NS	<i>ADH1A</i>
5	121739518	G	C	0.000352113	0.799286	0.403375	0.047536	NS	<i>SNCAIP</i>
5	140012292	G	A	0.000352113	0.799286	0.403375	0.047536	NS	<i>CD14</i>
5	149459791	G	A	0.000352113	0.799286	0.403375	0.047536	NS	<i>CSF1R</i>
5	174868811	G	A	0.000352113	0.799286	0.403375	0.047536	NS	<i>DRD1</i>
6	11190824	G	T	0.000352113	0.799286	0.403375	0.047536	NS	<i>NEDD9</i>
6	75953486	A	C	0.000352113	0.799286	0.403375	0.047536	NS	<i>COX7A2</i>
6	88854074	C	T	0.000352113	0.799286	0.403375	0.047536	NS	<i>CNR1</i>
6	160106042	G	A	0.000352113	0.799286	0.403375	0.047536	NS	<i>SOD2</i>
6	163735853	C	A	0.000352113	0.799286	0.403375	0.047536	NSS	<i>PACRG</i>
7	98257841	C	T	0.000352113	0.799286	0.403375	0.047536	NS	<i>NPTX2</i>
7	100490982	C	T	0.000352113	0.799286	0.403375	0.047536	NS	<i>ACHE</i>
7	140402694	C	T	0.000352113	0.799286	0.403375	0.047536	NS	<i>NDUFB2</i>
8	16850601	C	T	0.000352113	0.799286	0.403375	0.047536	NS	<i>FGF20</i>
8	18258061	A	C	0.000352113	0.799286	0.403375	0.047536	NS	<i>NAT2</i>
8	18258273	G	A	0.000352113	0.799286	0.403375	0.047536	NS	<i>NAT2</i>
8	20022397	G	C	0.000352113	0.799286	0.403375	0.047536	NS	<i>SLC18A1</i>

8	32463147	A	G	0.000352113	0.799286	0.403375	0.047536	NS	<i>NRG1</i>
8	32617779	T	G	0.000352113	0.799286	0.403375	0.047536	NS	<i>NRG1</i>
8	42262388	A	C	0.000352113	0.799286	0.403375	0.047536	NS	<i>VDAC3</i>
9	124906576	C	T	0.000352113	0.799286	0.403375	0.047536	NS	<i>NDUFA8</i>
10	102289129	T	C	0.000352113	0.799286	0.403375	0.047536	NSS	<i>NDUFB8</i>
10	102289200	G	A	0.000352113	0.799286	0.403375	0.047536	NS	<i>NDUFB8</i>
11	2189347	C	T	0.000352113	0.799286	0.403375	0.047536	NS	<i>TH</i>
11	67803768	A	C	0.000352113	0.799286	0.403375	0.047536	NS	<i>NDUFS8</i>
12	4763550	C	T	0.000352113	0.799286	0.403375	0.047536	NS	<i>NDUFA9</i>
12	9254254	C	T	0.000352113	0.799286	0.403375	0.047536	NS	<i>A2M</i>
12	40713845	G	C	0.000352113	0.799286	0.403375	0.047536	NS	<i>LRRK2</i>
12	99064865	G	A	0.000352113	0.799286	0.403375	0.047536	ESS	<i>APAF1</i>
14	64749426	C	T	0.000352113	0.799286	0.403375	0.047536	NS	<i>ESR2</i>
15	89867387	C	T	0.000352113	0.799286	0.403375	0.047536	NS	<i>POLG</i>
16	2132510	C	T	0.000352113	0.799286	0.403375	0.047536	NSS	<i>TSC2</i>
16	2134268	G	A	0.000352113	0.799286	0.403375	0.047536	NS	<i>TSC2</i>
16	2134547	G	A	0.000352113	0.799286	0.403375	0.047536	NS	<i>TSC2</i>
16	72108269	C	T	0.000352113	0.799286	0.403375	0.047536	NS	<i>HPR</i>
16	72110713	C	A	0.000352113	0.799286	0.403375	0.047536	SG	<i>HPR</i>
19	13397623	C	T	0.000352113	0.799286	0.403375	0.047536	NS	<i>CACNA1A</i>
19	14677646	C	T	0.000352113	0.799286	0.403375	0.047536	NS	<i>NDUFB7</i>
20	61981303	C	T	0.000352113	0.799286	0.403375	0.047536	NS	<i>CHRNA4</i>
22	35783080	C	T	0.000352113	0.799286	0.403375	0.047536	NS	<i>HMOX1</i>
11	2190899	C	T	0.000352609	0.797836	0.40338	0.0479429	NS	<i>TH</i>

**Table 24: A list of coding, protein-altering, highly significant single variants with p-values < 0.05**

Chromosomal location, reference and alternate alleles, effect size, p-value, effect type and corresponding gene are listed for each candidate variant. Abbreviations: NS = Non-synonymous. NSS = Normal splice site. SG = Stop gain. ESS = Essential splice site. CL = Codon loss.

### 3.3.5.3.2 Comparison with MSA GWAS results

Upon analyzing the gene burden test results and identifying some very significant findings, we compared this with the MSA GWA study results to determine any overlapping loci.<sup>323</sup> Our findings revealed that 7 loci identified by the nearest gene in the MSA GWA study data demonstrated significant single variant burdens in our exome cohort, two of which alter protein-coding, *XDH* and *CDH4* (Table 25). A full list of the most significant loci from the MSA GWA study is listed in Appendix section 8.1.5

CHR	POS	REF	ALT	ID	Pooled_Alt_AF	DIRECTION_BY_STUDY	EFFECT_SIZE	EFFECT_SIZE_SD	H2	PVALUE	Gene	Consequence	MAF
2	31564244	A	G	.	0.00314027	+	0.330135	0.140179	0.00387053	0.0185181	XDH	non-synonymous*	.
20	60503350	G	A	rs6142884	0.49162	+	0.0319617	0.0154777	0.00297784	0.0389221	CDH4	non-synonymous	57.20%

**Table 25: Single non-synonymous variants with p-values < 0.004 in top genes from MSA GWA study**

The MAF of the variant in *XDH* is unknown, while the MAF of the variant in *CDH4*, 57.20%, is quite common.

### 3.3.5.3.3 Comparison with WES results

After thoroughly reviewing the gene burden and single variant data, we searched for genes harboring single non-synonymous significant variants corresponding to those identified by the gene-based approach to WES analysis (3.3.4). While we did not identify any of the same variants confirmed by Sanger sequencing, we did find several non-synonymous variants with p-values <0.05 in 4 of our candidate genes: *CASP8AP2*, *RECQL4*, *RNF112*, *BCL2L13*.

#CHROM	POS	REF	ALT	N	POOLED_ALT_AF	EFFECT_SIZE	EFFECT_SIZE_SD	PVALUE	Effect	Gene
6	90575686	A	G	1420	0.000352113	0.799286	0.403375	0.047536	Nonsynonymous	CASP8AP2
6	90577876	G	T	1420	0.0028169	0.406989	0.146203	0.00537378	Nonsynonymous	CASP8AP2
8	145739024	C	T	1420	0.000352113	0.799286	0.403375	0.047536	Nonsynonymous	RECQL4
8	145741992	G	A	1420	0.000352113	0.799286	0.403375	0.047536	Nonsynonymous	RECQL4
17	19314726	C	G	1420	0.000352113	0.799286	0.403375	0.047536	Nonsynonymous	RNF112
17	19319333	A	G	1419	0.000704722	0.783217	0.294633	0.0078541	Nonsynonymous	RNF112
22	18209554	C	G	1420	0.000352113	0.799286	0.403375	0.047536	Nonsynonymous	BCL2L13

**Table 26: Significant non-synonymous single variants in genes identified by WES filtering pipelines**

Among all 4 genes, the only significant gene that overlapped with our RAREMETAL analysis was *CASP8AP2*.

### 3.4 Discussion

According to the PRL hypothesis, a combination of variants, ranging in both graded risk and frequency, unify the genetic etiology of complex disease. While the technology to investigate this hypothesis and identify variants can be approached from several dimensions, the methodology must be tailored according to both the disease under investigation and any prior genetic knowledge. It therefore follows that investigation of an incredibly rare and understudied disease like MSA with no known genetic etiology exemplifies a very challenging task. Further complicating the situation is the significant rate of misdiagnosis, which can create extensive bias in the results. However, given the rareness of the disease, the ability to acquire a sufficient number of only pathologically confirmed samples while obtaining requisite statistical power to detect associations is a tremendous obstacle in itself.

We undertook a series of experiments aimed to generate a list of genes as evidence based candidates for association with MSA. The intent of this work was to generate a list that could be published and that would allow others to attempt validation of these as genuine risk genes/variants.

The identification of variants through WES MAF filtering represented the first practical step towards the dissection of MSA genetic architecture in the context of rare

variants. Using the maximum number of samples we could feasibly obtain for MSA, we performed extensive MSA WES and analyses using several different approaches.

Notably, because of the inherent limitations of studying rare variability in this cohort, we designed this as a hypothesis generating study – where several analytical approaches were used to generate a list of candidate variants and genes that could be published so that others could independently replicate these. We feel that this approach is ultimately the most efficient method to understand the genetic architecture of disease.

Following quality control analyses in the locally processed case/control cohort, we searched for any PD associated genes that had been identified in familial and sporadic studies, at this point of the analysis no mutations in PD-linked genes were identified.

Next, we moved on to our first analytical approach by identifying shared rare variants among the MSA sporadic case cohort. While this is an ideal pipeline when analyzing families or relatives affected with disease, as one would expect the same mutation to be derived from a common ancestor, this is not always the case for sporadic cohort analyses. Further, as do not know the pattern of inheritance, we included all heterozygous, homozygous and compound heterozygous alleles in our candidate variant lists.

As this was our first approach towards analysis, we decided it was most logical to apply harsh filters and generate a small candidate list, all of which could all be confirmed or rejected by Sanger sequencing. If we could not identify a variant, we could expand this list and modify our filters to be more inclusive. This seemed more rational than initially using lenient filters, which would yield an impractically large list of candidate variants that would make analysis even more complicated.



Among all of our candidate variants, several appeared promising regarding function (i.e. *DNAJC11* is also a member of the gene family including *DNAJC6*, which harbors a homozygous variant known to cause autosomal recessive forms of atypical PD). Likewise, a novel variant identified by the VCF in *VCP* was quite interesting, as variants in *VCP* are known to cause familial forms of ALS. As our criteria required at least 2.5% of samples in the MSA cohort to carry the variant, which was approximately 9 or more individuals, we tested at least 5 different MSA individuals to determine the authenticity of each candidate variant. If any pathologically confirmed samples were called as heterozygous for the variant, they were prioritized and tested first. Despite extensive primer design and sanger sequencing, we failed to confirm these variants.

Before moving onto a more liberal filtering approach, we acknowledged that many of our previous candidates were indels or frameshift mutations, and that there is a well documented and very high false positive rate with this form of variation called by NGS. It was important to acknowledge the limitations of WES, which has a tendency to misalign and incorrectly call variants longer than SNVs. Cognizant of this sequencing bias, we decided to focus on novel SNVs in the form of non-synonymous SNPs, stop gain, and stop loss variants, as we believed the probability of SNV validity was considerably higher.

Using our gene based filtering approach, we not only incorporated less stringent filters but also defined more lenient criteria for candidacy inclusion, whereby at least 1% of all MSA cohort individuals must share any novel variant in the same gene. While >1% of the cohort (at least 4 individuals) is small indeed, our focus on solely analyzing novel SNVs gave us confidence in this approach.

After significantly distilling our list, we pursued Sanger sequencing of all variants with sufficient DNA, confirming 75% (27) of the tested 36 variants. As we were completely unbiased regarding gene function (if known) and expression in the brain, we eliminated variants that confirmed but were predicted as benign polymorphisms from our candidate list. However, as some variants did not confirm in a single gene but other variants in that gene could not be tested due to insufficient sample DNA, we kept those genes in our candidate list for upcoming burden analyses.

The next analysis centered on using the RAREMETAL R-package to investigate gene burden and individual variant analyses. To execute this work we used a novel workflow developed by LNG within googleCloud. Cloud based analysis is not only advantageous regarding the speed of analysis, but the alignment and sensitivity of variant detection are remarkably improved from local data processing. Notably we believed that the ability to align and call variants in exome data from cases and controls of different provenance in parallel was likely to improve the sensitivity of variant detection and reduce errors due to batch effects. Notably, we identified a large number of variants that were not apparent in our previous analyses on this data set, including several highly relevant to neurodegenerative disease. A highly significant gene burden signal at *LRRK2* was unexpected based on the WES locally processed and filtered results. Likewise, several other genes that were not previously identified required further scrutiny by ascertaining all non-synonymous variants in genes with strong neuronal expression. Based on our “in-depth gene analysis,” several non-synonymous variants appeared to exist exclusively in cases or were present in a disproportionally higher percentage of cases than controls. Sanger sequencing of variants in *SLC44A5* and *GLIPRI* failed to

confirm with sanger sequencing, while the variant rs150022229 in *CASP8AP2* was validated in all 5 MSA samples tested from different cohorts (French, UK, USA). While we can only state that 25% (1 of 4) of our Sanger sequenced variants from genes derived exclusively from the Googlegenome pipeline were validated, we acknowledge several caveats to this statement: first, our selection bias of variants to pursue, and second, that  $n=4$  is not indicative of authenticity among the entire pipeline derived results. As we must interpret these preliminary results with caution and recognize inconsistencies in Sanger sequencing validation of results, the next most practical step is to perform a combined validation replication set. As opposed to individually testing each significant and functionally relevant variant, we should seek to combine both validation and replication by using an entirely new MSA cohort. While we recognize that sample acquisition is a significant obstacle, even a small cohort will shed light on authentic associations, paving the groundwork for future genetics and functional work.

Among all the results, perhaps the most interesting finding was the presence of *LRK2* p.G2385R in 10 MSA cases, two of which are homozygous. However, disappointingly we were unable to confirm this variant with Sanger sequencing. Although the genotypes identified by local and GoogleGenomics pipelines were 99% concordant (M. Nalls and R. Gibbs personal communication) there are rare differences. Some of these differences are artifactual variant calls that one pipeline may be more prone to call and these are often the basis of some of the more extreme association signals; this appears to be the case with p.G2385R and once again illustrates the need for validation/replication.

Based on the results discussed thus far, there are a few genes which should be prioritized in the combined validation replication set to unravel MSA genetic architecture (Table 27).

GENE	P-value: gene burden	P-value: single variant
<i>LRRK2</i>	4.05E-07	2.47E-06
<i>SLC44A5</i>	9.76E-10	1.26E-12
<i>GLIPR1</i>	4.74E-09	1.94E-09
<i>CASP8AP2</i>	0.0131902	0.00537378

**Table 27: Most significant and functionally relevant genes identified in hypothesis generating dataset.**

Each gene harbored several significant variants but those with the lowest p-values are listed above.

When comparing the results of samples using local alignment vs. GoogleGenome, one of the genes identified in the gene based filtering approach of local WES analysis harbored significant variants in the burden analyses. While none of these variants in *CASP8AP2* had a p-value <  $1 \times 10^{-6}$ , all were below the standard 0.05 cut-off. *CASP8AP2*, also known as *FLASH*, is known to play a key role in several cellular processes such as apoptosis regulation, mRNA processing and influencing gene expression via transcriptional regulation. Further, research has suggested that FLASH protein is a component of the death-inducing signaling complex that includes the Fas receptor, Fas-binding adapter FADD, and caspase 8, while maintaining a regulatory role in Fas-mediated apoptosis (<http://www.genecards.org>). Recent work has also revealed a role in proteasome-dependent degradation.<sup>344</sup> As apoptosis and ubiquitination are important processes that oligodendroglia undergo in MSA pathophysiology, the role of *CASP8AP2* as a plausible risk factor in MSA development is intriguing. Our case frequency (6/736) of 0.82% vs. our control frequency (3/2229) of 0.13% suggests that independent replication of MSA samples harboring a *CASP8AP2* gene burden will be required to

confirm any association, but these preliminary results set the stage for the replication phase (Figure 33).

In our single variant burden analyses, several significant variants were identified in PD associated genes like *LRRK2* and *PARK2*. By running a functional annotation analysis on DAVID, there were many single variants harboring significance with a p-value  $< 1 \times 10^{-6}$ , making ideal candidates for further investigation.

Comparison of the results generated via Googlecloud and RAREMETAL with the top loci or genes from the MSA GWA study was informative, as two single non-synonymous variants in genes identified by GWA data demonstrated significant single variant burdens. While the MAF of the allele in *CDH4* is extremely common, this suggests a possible role in graded risk for MSA. As GWA studies are targeted toward the identification of common risk variants, the identification of *CDH4* through this methodology seems plausible. The second gene, *XDH*, harbors a novel non-synonymous variant with a significant burden in the MSA WES dataset run through the Googlegenome pipeline. As this is a novel variant, we would anticipate a very low MAF, with possibly damaging effects. Resonating with our previous hypothesis generated results from burden analyses, the significant variants identified in *CDH4* and *XDH* should be likewise prioritized in future independent replication cohort analyses.

In addition to shared genes and/or variants between the Googlegenome single variant burden data and MSA GWA study results, there was substantial overlap in single variants within several genes identified through WES local alignment analyses. While *CASP8AP2* also demonstrated a significant gene burden, single variants in *RECQL4*, *RNF112* and *BCL2L13* were also statistically significant. *RNF112*, is an interesting candidate, as it is

primarily expressed in the brain and known as the “ring finger” protein, it plays a critical role in neuronal and glial cell differentiation.<sup>345</sup> Further, while variants in *RNF112* are known to cause Smith-Magenis syndrome, there is close association with some forms of SCAs and ataxias.<sup>345</sup> While we acknowledge that our hypothesis generated results are preliminary, the functional relevance likewise makes some of these genes attractive candidates to pursue.

While there was clear overlap in some of our candidate genes between local and GoogleCloud pipelines, the majority of genes identified by RAREMETAL analyses on GoogleGenome were new. Most notably were the multiple variants revealed in *LRRK2*, with p.G2385R being the most striking. However, this is an example of a false positive association, being driven by an artifact of the variant calling process, and illustrates the need for validation and replication. The ability to perform both alignment and processing of data locally and on GoogleGenome was extremely insightful towards our analytical interpretation of the results, as we were able to achieve a comprehensive outlook of all possible significant variations and associations. The difference in overlap sheds light on the variation in data quality and sensitivity between both pipelines. While this is not only advantageous by providing the greatest number of results to analyze for a substantial hypothesis generating dataset, it will also play a pivotal role in guiding data interpretation of future WES projects and the execution of WES association analyses.

As there are inherent limitations to WES regarding capture rate and coverage, our ability to achieve high quality depth and coverage is a testament of these limitations using our local pipeline. Thus, as we were afforded the opportunity to incorporate this same

dataset on GoogleCloud, we can directly compare the differences and recognize the superiority of the latter regarding scope of detection.

Specifically, RAREMETAL offers some very unique advantages that cannot be performed through local alignment. Fundamentally, this entails the reconstruction of gene-level statistics from single variant score statistics through the generation of several unique reports for each gene-level test. Further, RAREMETAL is able to generate single variant and gene burden information in visually aesthetic graphics including QQ and Manhattan plots.<sup>346</sup> Moreover, the ability to create unique graphs based on desired MAF level is also a very effective tool for rare variant analysis in a hypothesis generating dataset. Finally, the several types of burden tests incorporated into the RAREMETAL analysis package, including the SKAT, MB, VT and CMC allow the user to determine which test is the most appropriate for a particular study; hence, as we were uncertain about the magnitude and direction of causality of variants within a single gene, we were able to generate results of non-burden (SKAT), weighted aggregation (MB), adaptive burden (VT), and combined burden (CMC) tests to assess which gene-level statistics attained substantial power, as reflected by p-values.

While the advantages are numerous, perhaps the largest disadvantage to using RAREMETAL and Googlegenome was the miscalling of some variants with very significant p-values (i.e. *LRRK2* p.G2385R). While miscalling is an inherent feature of WES, WGS and other state-of-the-art technologies, 3 of the 4 variants of interest that we attempted to validate were artifacts. While we cannot draw conclusions on such a small sample size under scrutiny, recognizing this limitation is important for future studies nonetheless. That being said, performing the local analysis was a key first step, as we

generated a preliminary candidate gene list and even replicated some of these findings using the GoogleGenome pipeline. Thus, generation of results from both pipelines has allowed us to integrate disparate information into a more cohesive framework for future investigations.

As our goal was to obtain a hypothesis generated dataset, we have made progress in the discovery phase by identifying several candidate genes and variants. Ultimately, the next steps involve the acquisition of an independent MSA cohort for a combined validation replication investigation. If particular variants can be successfully confirmed in a subsequent cohort, a resequencing approach covering these genes would be valuable towards the identification of rare variants and possible SVs. As the resequencing protocol is applicable over diverse genomic regions including both small and large exons, short and long contiguous genomic targets, genome targets within repeats and even non-coding DNA, the genetic architecture of these genes can be further dissected.

As a very rare disease, we recognize the limitations in sample collection and the ability to seek pathological confirmation of disease. However, we believe we have overcome these limitations to our greatest abilities through acquisition of several hundred samples and comprehensive WES analyses using both local and GoogleCloud processing pipelines. As we have generated several candidates in this discovery phase, we have planted the seeds for future investigation of the genetic architecture of MSA in the scientific community.





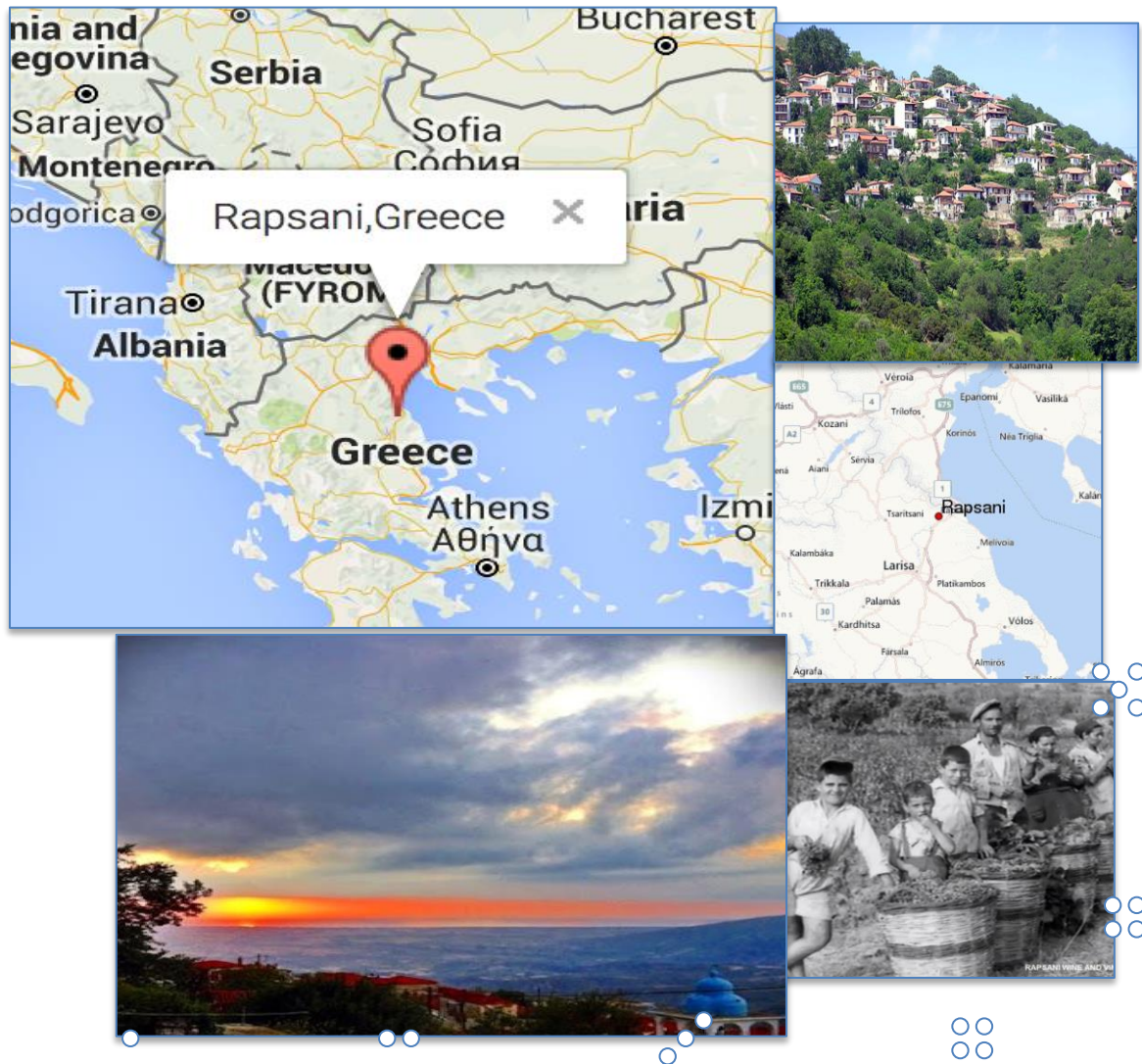
## 4 Exploring the genetic etiology of PD in the Greek village of Rapsani

***Statement of Contributions:*** Neurological examinations were administered by Dr.

Georgia Xiomerisiou and Dr. Henry Houlden. Clinical history, pedigree information and examination results were obtained by Dr. Georgia Xiomerisiou. Samples were received at both UCL and the Laboratory of Neurogenetics, NIA, NIH. I performed exome sequencing and analysis on all samples. I manipulated, filtered and annotated all files and determined several candidate gene lists and analyzed the whole genome genotype and whole genome sequencing data. I supervised the assessment of candidate genes and the confirmation of segregating variants.

### 4.1 Introduction

In recent history, the ease of global migration has facilitated assimilation and interbreeding of genetically heterogeneous individuals. Despite this extensive global diaspora, a handful of populations have remained fairly isolated. Such populations are ideal to study genetically, as the ability to control for normal ethnic variation is much more precise and such a population structure may result in the over-representation of a particular trait and a reduction in genetic and allelic heterogeneity. We have been afforded the privilege to collaborate with a physician in Greece, Dr. Georgia Xiomerisiou, who has carefully studied several Rapsani village members (~1500 individuals total) in the foothills of Mount Olympos, Greece (Figure 48).



**Figure 48: Rapsani, Greece.**

The village of Rapsani, embodying the symbolic Hellenic spirit and freedom, represents a culturally rich population, garnering fame in ancient arts, literature and education. Represented by historic landmarks such as the church and watermills, the village originated almost a millennium ago from the Byzantine era and flourished in wine production and viticulture. Traditionally, the Rapsani villagers have passed down the oral tale explaining that the village was created upon unification of four smaller villages in the

10<sup>th</sup> century. Interestingly, the village achieved such extreme isolation due to the migratory restrictions during the Turkish reign in Greece from 1455-1821, as Turkish individuals were forbidden to inhabit this territory. The only interbreeding that occurred during this time is thought to be with other indigenous individuals, proximally located in central Greece, seeking refuge in order to escape Turkish control. As historical records have not mentioned the occurrence of any natural disasters or other catastrophic events that would radically alter the genetic pool, the Rapsani village population has been considered to be stable since approximately 300 BC (personal communication G. Xiromerisiou).

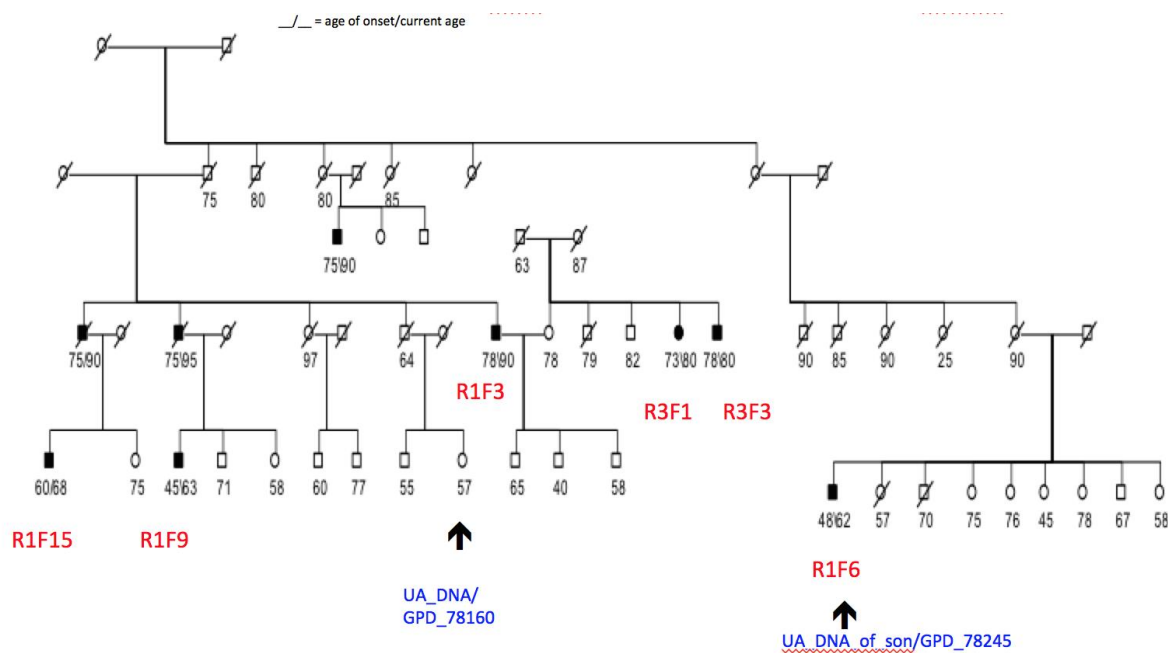
As any population isolated for sustained periods of time, the gene pool becomes extremely homogenous and the prevalence of individual mutations may be significantly higher than any other population (if it even exists elsewhere). Notably, the prevalence of PD among industrialized nations is approximately 0.3% among individuals of all ages, rising to 1% among those over 60 years of age and up to 4% among individuals over 80 years of age.<sup>63</sup> Among the 1500 registered members of the Rapsani village and 600 permanent residents, the estimated prevalence of PD in the general population is between 1-2%; however, as this value has not been reported in the literature, and is rather based upon communication with collaborators, we must interpret this with caution. Cognizant that this estimate is higher than the global prevalence, similar to the highly inbred Amish community with an estimated prevalence of nearly 1%, this suggests a possible genetic etiology of PD among the Rapsani village community (Northwest Parkinson's Foundation 2013, personal communication G. Xiromerisiou). After careful inspection, the pedigrees suggest that PD within Rapsani appears to be of a familial form, with

several nuclear families representing the majority of the burden of disease. With the help of Dr. Georgia Xiromerisiou, we have been able to obtain genetic samples from a few families within the village to embark on a comprehensive genetic analysis. In an effort to identify the genetic lesion(s) underlying PD within Rapsani village members, we have pursued whole genome genotyping, WES and WGS followed by extensive data analyses.

## **4.2 Materials and methods**

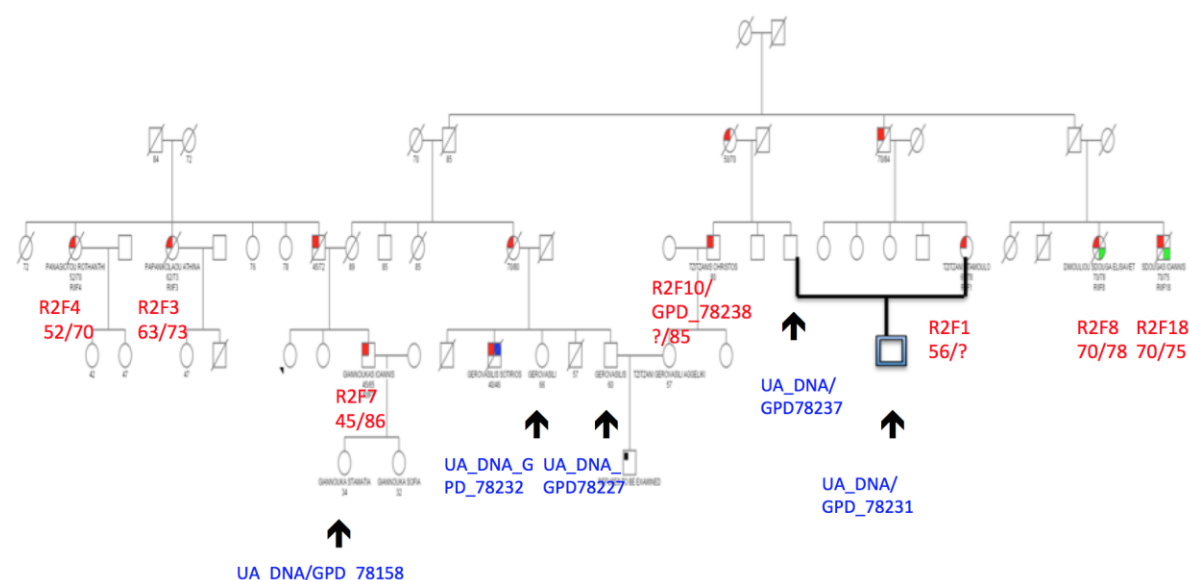
### **4.2.1 Subjects**

This study focused on 5 families from the Rapsani village. The first 3 (Rapsani I-III) have multiple family members affected with PD in a single pedigree. The remaining 2 families (Rapsani IV-V) are small families with 2 affected individuals (with gDNA from one member of each family). Recent consanguinity was not initially reported in any of these families. To further assist our analyses, we were able to obtain gDNA samples from several unaffected family members. One caveat, however, is that many of these individuals are relatively young and may develop PD later in life. Keeping this in mind, the following families were investigated (Figure 49, Figure 50, Figure 51).



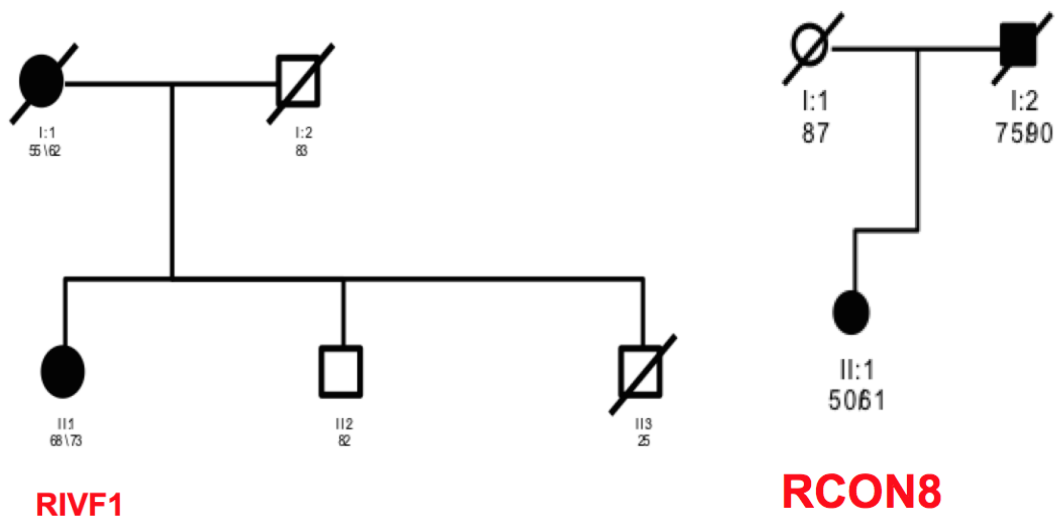
**Figure 49: Pedigree of Rapsani families I and III.**

‘/\_’ denotes age of ‘onset/current age’ (if known). All members shaded in black are affected with PD and those in white are unaffected. ‘R\_F\_’ represent sample identifications. All IDs in red text are affected individuals whose gDNA we were able to obtain. ‘UA\_DNA’ in blue text are unaffected individuals whose gDNA we were able to obtain. ‘UA\_DNA of son’ refers to the son of R1F6 who is not located on the pedigree, as we were unable to obtain more information about this nuclear family. Those with a diagonal slash are deceased.



**Figure 50: Pedigree of Rapsani family II.**

'\_/\_ ' denotes age of 'onset/current age' (if known). All members shaded in red in the top left quadrant are affected with PD and those with a white top left quadrant are unaffected. Other colors denote co-maladies. Those with green in the bottom right quadrant reflect individuals with dementia. Individuals with blue in the top right quadrant reflect those with motor neuron disease (MND). 'R\_F\_' represent sample identifications. All IDs in red text are affected individuals whose gDNA we were able to obtain. 'UA\_DNA' in blue text are unaffected individuals whose gDNA we were able to obtain. Those with a diagonal slash are deceased.



**Figure 51: Pedigrees of Rapsani families IV and V**

‘\_/\_’ denotes age of ‘onset/current age’ (if known). All members shaded in black are affected with PD and those in white are unaffected. All IDs in red text are affected individuals whose gDNA we were able to obtain. Those with a diagonal slash are deceased.

For the majority of affected individuals, a thorough clinical history was obtained by neurologists, Dr. Georgia Xiromerisiou and Dr. Henry Houlden. This involved the administration of several standard neurological tests with the results shown below in Table 28.

Clinical Phenotype	R1F3	R1F6	R1F9	R1F15	R3F1	R3F3	R2F1	R2F3	R2F4	R2F7	R2F8	R2F10	R2F18
Age of onset	78	48	45	60	73	78	65	63	52	45	70	?	70
Bradykinesia	✓	✓	✓	✓	✓	✓		✓	✓	✓	✓	✓	✓
Asymmetric Tremor (initially)	✓	✓		✓	✓	✓	✓	✓	✓		✓	✓	
Bilateral Tremor (initially or progressed)			✓				✓	✓	✓	✓	✓	✓	✓
Tremor: Upper Limb only	✓	✓	✓	✓			✓	✓	✓		✓	✓	✓
Rigidity	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
Hallucinations							✓				✓		✓
Dyskinesias			✓				✓	✓	✓	✓	✓		✓
Response to Levodopa: primary	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	
Response to Levodopa: sustained	✓		*		*							✓	
Motor complications		✓	✓	✓	✓	✓		✓	✓	✓	✓		✓
Sleep disturbances (Insomnia)							✓						
Dementia		✓					✓				✓		✓
Autonomic Dysfunction					✓	✓	✓						
Hypophonia	✓	✓			✓	✓							
Stooped posture		✓			✓	✓	✓						

**Table 28: Clinical Phenotypes of Rapsani families I-III.**

\*Sustained Response to Levodopa: R1F15 experienced a sustained response to levodopa for several years but recently experiences wearing off periods. R1F9 has impulse control disorder and has had to discontinue using any other dopaminergic agonists despite a very positive response.

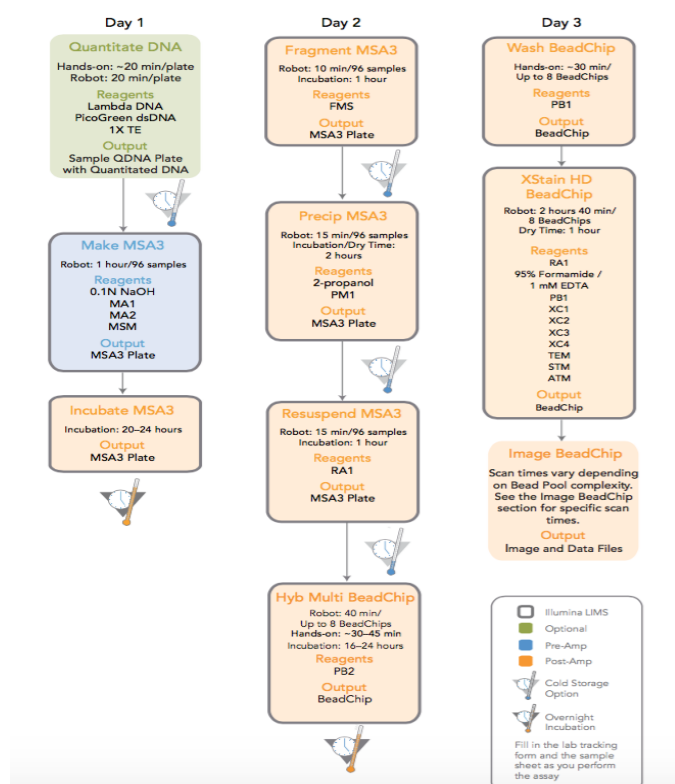
\*\*Autonomic dysfunction includes severe constipation and urinary disturbances.

Clinical info could not be obtained for individuals from Rapsani families IV (R1VF1) and V (RCON8).



## 4.2.2 Genotyping

All samples were genotyped according to the manufacturer's instructions on the OmniExp-12,v1.0 DNA Analysis BeadChip (Illumina Inc., San Diego, CA). The protocol workflow is illustrated in Figure 52.



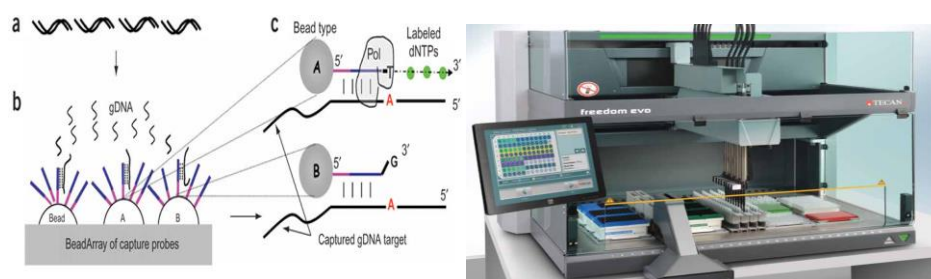
**Figure 52: Illumina Infinium Genotyping Workflow.**

Core steps for sample processing include: MSA3 generation, incubation, fragmentation, precipitation and resuspension. This is followed by beadchip hybridization, washing, Xstaining and imaging.

Reproduced by ([http://support.illumina.com/array/array\\_kits/humanomniexpress-12-beadchip-kit/documentation.html](http://support.illumina.com/array/array_kits/humanomniexpress-12-beadchip-kit/documentation.html)).

Genotyping requires 400 ng of gDNA per sample in a volume of 8 ul. First, each sample is denatured in 0.1N NaOH and amplified at 37°C for 20–24 hours. Next, the

denatured DNA is enzymatically sheared at 37 °C for 1 hour to create fragments of around 300bp. This is followed by precipitation using 2-propanol and re-suspending in re-suspend amp1 solution (RA1) (Illumina, San Diego, CA, USA). Next, the DNA fragments are denatured at 95 °C for 20 minutes, dispensed upon the BeadChips using a robot, and hybridised for 16-20 hours at 48 °C. Following incubation, the BeadChips are washed in order to remove redundant DNA fragments. This is followed by automated allele-specific extension and staining reaction using a Tecan Freedom EVO robot (www.tecan.com, Männedorf, Switzerland). Next, a second round of washing using PB1 solution (Illumina), vacuum-drying (1h) and finally imaging of the BeadChips using a two-color confocal laser system in a BeadArray Reader (Illumina). A workflow of the overall genotyping process is demonstrated in Figure 53.



**Figure 53: Overall genotyping workflow**

- A. Fragmentation of DNA. B. Hybridization of DNA fragments to Beadchip capture probes. C. Allele-specific staining using dNTPs and polymerase. D. Imaging of Beadchips using two-color confocal laser system on the Tecan. Modified version reproduced from (Gunderson et al., 2005) and ([www.tecan.com](http://www.tecan.com)).

Data resulting from array scanning is loaded into Genome studio (Illumina Inc, CA). According to default parameters, samples self-cluster for each SNP from the Beadchip and are then merged to form a cohort project file. Once self-clustering is complete, a final report is generated that is used to form .map and .ped files for input into PLINK.

### 4.2.3 Quality control

Stringent QC analyses are necessary to ensure removal of samples and SNPs that could bias results. QC and statistical analyses were performed using either Plink (Purcell, et. al., 2007), GERMLINE (<http://www1.cs.columbia.edu/~gusev/germline/>) and R (<https://www.r-project.org/>) in a Linux Ubuntu system. Quality control for each sample involved removal of individuals with the following: missing or misreported gender (--missing gender), <95% call rate (--geno), SNPs with MAF <0.1, SNPs on the X-chromosome with MAF <0.05, Hardy-Weinberg Equilibrium (HWE) ( $p < 10^{-5}$ ) standard deviations from the mean, and >10% heterozygosity rate. As a standard quality control check, we performed a multidimensional scaling (MDS) analysis with PLINK (using the --cluster --mds-plot 4 commands) based on the first four principal components to confirm uniform European ancestry.

### 4.2.4 Identifying runs of homozygosity

Using the .map, .ped and .fam files generated from the genotyping final reports, we had the ability to detect long homozygous segments in the data. This required a series of command line instructions, including making a binary bed file (--make-bed), which comprises all the alleles for subjects at every SNP site genotyped. A plink.hom file can be generated (using the --homozyg) command to create a table of all long homozygous

segments in each individual, with the corresponding genomic address and SNP ID. Next, a `plink.hom.overlap` file was generated (using the `--homozyg--group` command) to obtain regions of overlapping and potentially matching segments. This provides information about the number of segments in the cohort that match one another and the allelic match grouping of each segment, which can be compared with phenotypic status. Finally, to obtain all of the individual genotypes of overlapping segments, the `--verbose` command is used, generating an expanded version of `plink.hom.overlap` file.

In addition we performed homozygosity analysis outside of the Plink framework using homozygosity mapper (<http://homozygositymapper.org/HomozygosityMapper/index.html>). The input information can exist in VCF format, based on WES or WGS data, or as a genotyping final report. By assigning an affected or unaffected (control) status to each sample, the corresponding genotype information may be incorporated into a Manhattan plot, allowing comparison of regions of homozygosity along each chromosome. Once the data is uploaded, projects can be re-analyzed by the addition or subtraction of certain individuals. With the large quantities of data generated by the former three methodologies, we were able to assess homozygosity based on the following: common variants (obtained by genotyping), rare coding variants in the form of SNPs and short indels (obtained by WES), and CNVs and SVs for seven affected samples (obtained by WGS).

#### **4.2.5 Identifying segments identical by descent**

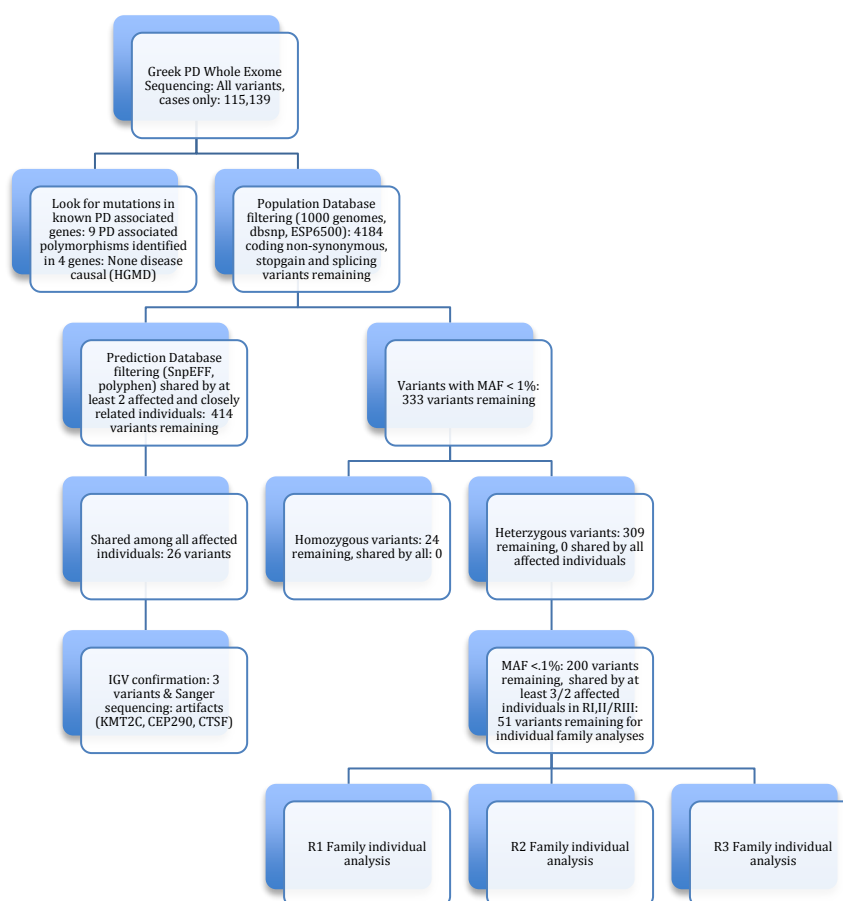
In order to detect shared segments among individuals of common ancestry, we used the Identity by Descent (IBD) segmental sharing option on Plink. Using familial

samples, this involves a series of steps: first, a plink.genome file is created (using the --genome command) to ensure a homogenous sample set. Next, the SNP set must be pruned in order to remove SNPs in LD, as the segmental sharing analysis requires using SNPs in linkage equilibrium. This will create a plink.prune.in file, which can then be transformed into .fam and .bim files (using the --make bed command). Once these files are generated, shared segments are determined through the --segment command and the desired length or number of SNPs included can be specified by additional commands. Using GERMLINE, we were able to generate this information by uploading the .ped and .map files of the Greek PD Rapsani Cohort. If one can identify long shared segments between distantly related affected individuals (i.e. from different Rapsani families), this may reflect IBD due to a common ancestry (i.e. founder mutation), as opposed to Identity by State (IBS). Hence, by identifying regions that are truly IBD among distantly related affected individuals, these regions are likely to harbor disease-associated etiology.

#### **4.2.6 Whole exome sequencing**

Whole exome sequencing was performed on all 23 samples obtained, including 16 affected and 7 unaffected individuals. The Illumina Nextera protocol was used which requires 50ng of gDNA per sample. The step-by-step details are explained in section 3.2.4. The entire wet-lab procedure including DNA library preparation and enrichment, clustering on the C-bot, and parallel sequencing by synthesis on the Illumina Hi Seq 2000 is the same. Regarding raw data analysis, the first several steps are also identical, including: mapping, alignment and duplicate removal, followed by raw variant callings and file conversions, incorporation of reference databases, assignment of quality (Phredd) scores to all variant calls, and generation of group VCFs (gVCFs).

The next steps involved downstream analysis and filtering of the gVCF. As this is a familial segregation analysis (as opposed to an association analysis), we are looking for shared variants between individuals. Thus, a gene-based candidate approach would not be a practical since our primary hypothesis was that disease is driven by a single shared variant among all affected individuals in the same gene. In order to identify such a variant, which would be rare a filtering approach was employed. A variant filtering pipeline for the Greek PD Rapsani village members is illustrated below (Figure 54). As a first step, we wished to exclude causal mutations in known PD-linked genes. Once this was shown to be negative, variants were filtered based on retaining variants with a frequency of <1%; filtering was performed against 1000 genomes, dbSNP and ESP6500 databases. Both heterozygous and homozygous variants were then prioritized based on segregation within all affected individuals in Rapsani and within individual nuclear families (Figure 29).



**Figure 54: Whole exome sequencing variant filtering pipeline for Greek Rapsani PD cohort**

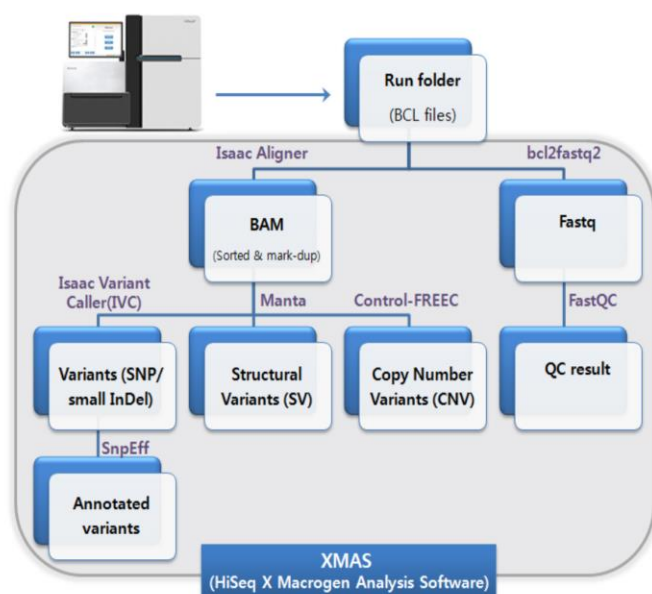
After looking for mutations in known PD associated genes, we used population database filtering followed by either prediction database filtering or MAF exclusion criteria.

#### 4.2.7 Whole genome sequencing

Given the limitations of exome sequencing, including capture efficiency among others, we decided to pursue WGS for a more comprehensive analysis. In the interests of efficiency, we outsourced WGS to the biotechnology company, MacroGen (<https://www.macrogenusa.com/>), using the Illumina TruSeq DNA PCR-free library preparation protocol. Seven samples were sent to MacroGen. This consisted of seven

affected individuals among Rapsani families I and II. From Rapsani family I, this included: R1F3, R1F6, R1F9 and R1F15. From the Rapsani II family, this included: R2F3, R2F4 and R2F7.

After each sample passed quality control filters and gender checks, the results were obtained in the form of fastq, BAM and individual VCFs for each sample. As WGS has the ability to capture long structural variants and CNVs, we obtained information according to the following workflow used by MacroGen (Figure 55).



**Figure 55: MacroGen whole genome sequencing analytical workflow**

WGS analysis uses analogous raw variant calling and alignment methods to WES but also requires additional programs to analyze SVs and CNVs using Manta and Control-FREEEC, respectively.

After BAM files are created using Isaac Aligner (Raczy et al 2013), specific programs are used to identify SNPs, Structural Variants (SV) and CNVs. SNPs and small indels, which are detectable in WES in coding regions only, are called using Isaac Variant Caller (IVC)



([www.illumina.com/documents/products/whitepapers/whitepaper\\_isaac\\_workflow.pdf](http://www.illumina.com/documents/products/whitepapers/whitepaper_isaac_workflow.pdf))

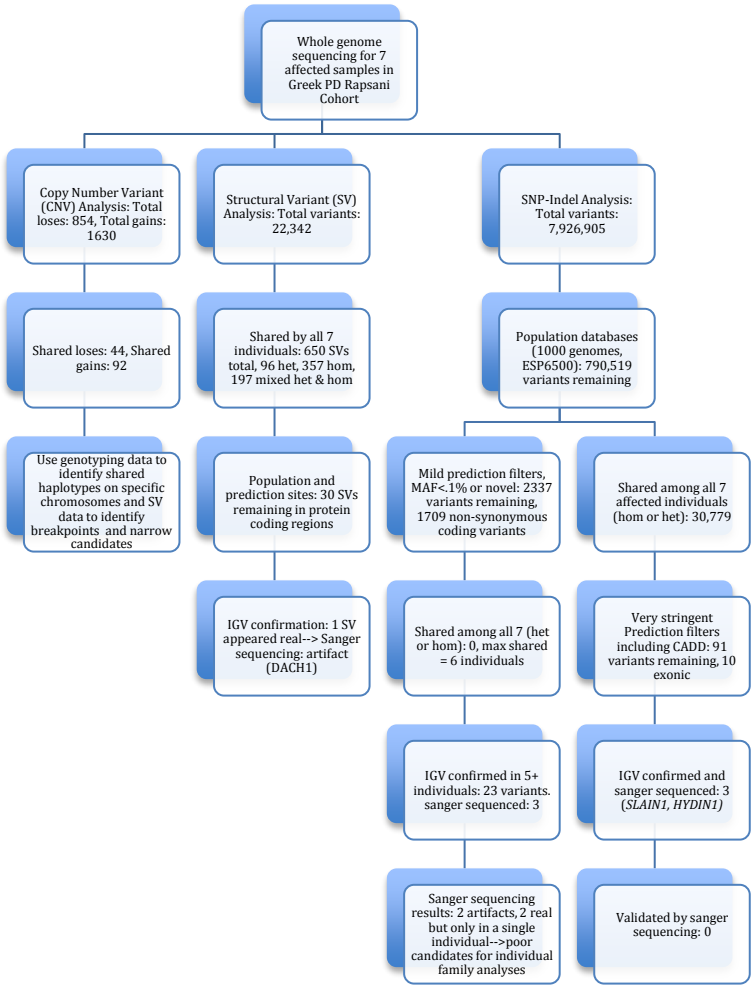
during WGS data analysis. SVs are called using Manta, which does not require split reads or successful breakpoint assemblies to accurately report a variant

(<https://github.com/StructuralVariants/manta>). CNVs are called using a program called Control-FREEC, which automatically calculates copy number and allelic content profiles and uses this information to predict genomic modifications including gains and losses (Boeva et al 2012). Lastly, SNPs are annotated using SnpEff using the following annotation pipeline steps: (1) Gene annotation based on hg19 coordinates (2) dbSNP138 ID mapping (3) dbSNP142 ID mapping (4) 1000 Genomes phase I release v3 mapping (5) ESP6500 data mapping (<http://snpeff.sourceforge.net/SnpEff.html>)

As we are looking for a single shared variant among the cohort, we merged the individual VCFs into a gVCF and followed the variant filtering pipeline illustrated below (Figure 56).

The WGS analyses were divided into several separate analyses according to the type of variant information obtained from sequencing. First, we pursued preliminary analyses of CNV data by determining all shared CNVs (both lost and gained) by all 7 affected Rapsani PD members. Secondly, we looked at SVs that were not captured using WES due to their size. Likewise, we assessed shared SVs among all affected members and further distilled our candidate SV list using population and prediction filters. Thirdly, we performed an extensive SNP-Indel analysis using two distinct approaches; the former was significantly more stringent, requiring the use of harsh prediction filters that left us with very few exonic variant candidates. The latter approach, however, was much more

lenient, using mild prediction filters in an effort to maintain all possible candidate variants.



**Figure 56: Whole genome sequencing variant filtering pipeline for Greek Rapsani PD cohort**

WGS investigation was focused on three types of variant analyses: SVs, CNVs and SNP-Indel variants.

#### 4.2.8 *C9ORF72* hexanucleotide repeat screening

To confirm that all Greek Rapsani individuals did not carry the GC rich, intronic repeat region in *C9ORF72*, which is a cause of familial ALS and FTD and which would have been difficult to detect using WES, we performed the repeat-primed PCR protocol.

This entails a PCR master mix with several reagents and a separate Repeat-Primer master mix listed below in Table 29 and Table 30.

Reagent	Volume in a 28ul Reaction (ul)	Final Amount
Roche Faststart Mix	14	1X
Primer Mix	2	0.7-1.4uM
DMSO	2	7%
Q Solution	5	1X
Deaza dGTP (5mM)	1	0.18mM
MgCL2 (25mM)	1	0.9mM
DNA (50ng/ul)	2	100ng

**Table 29: PCR master mix used for *C9ORF72* screening in all Greek samples**

Primer Mix	Volume in 100ul reaction (ul)	Final Amount
F1 (100uM)	20	20uM
R(100uM)	10	10uM
Tail (100uM)	20	20uM
H2O	50	N/A

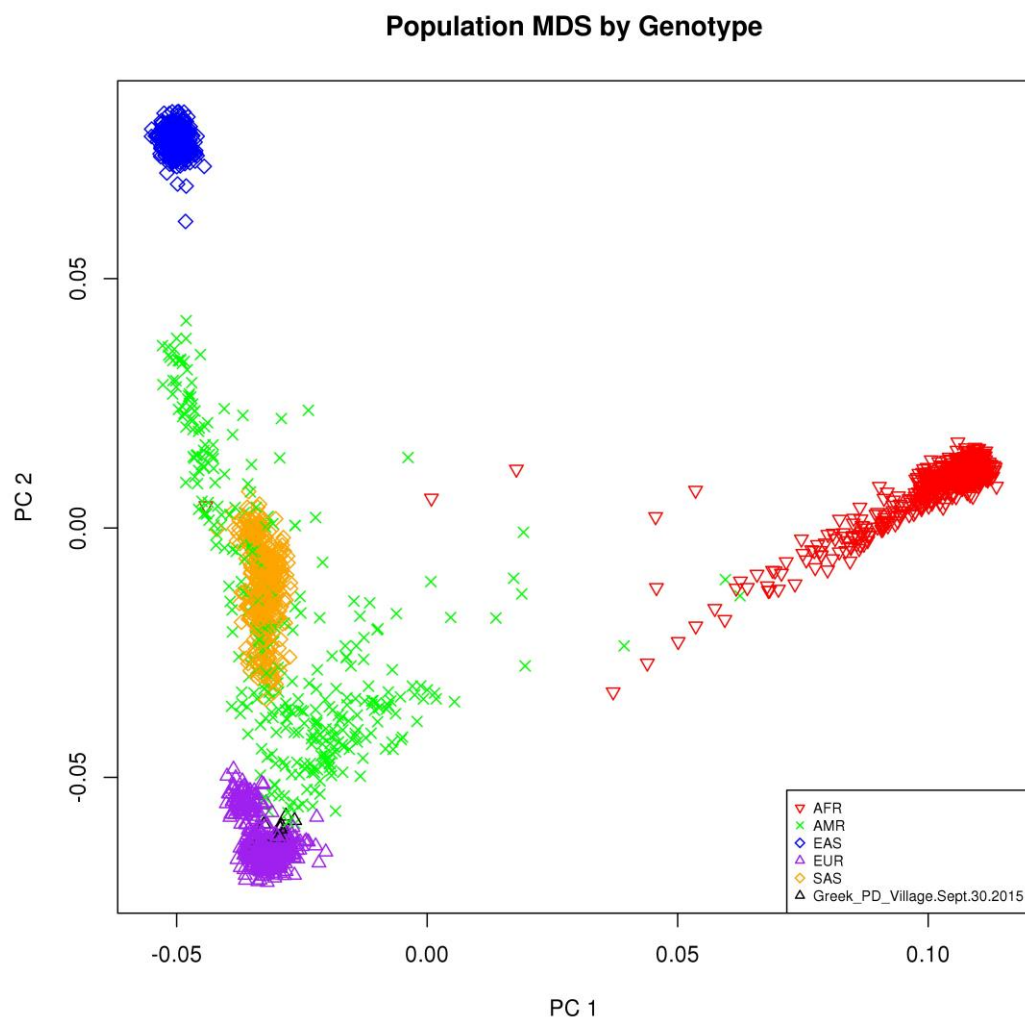
**Table 30: Repeat primer mix used for *C9ORF72* screening in all Greek samples**

The Repeat Primer sequences and thermal cycler conditions for this protocol are listed in section 8.1.2.5-8.1.3. After PCR amplification samples (in a 96-well plate) are heated to 95°C for 3 minutes and then placed immediately on ice for 5 minutes. The samples are then run on the ABI 3730 (Applied Biosystems, CA, USA) and the resulting data is analyzed using the GeneMapper program (ThermoFisher scientific: Life Technologies, Frederick, MD, USA). PCR amplification requires 3 reagents: 2ul PCR product, 0.5ul Liz500 size standard (ABI 3730), and 7.5ul HiDi formamide (ABI 3730). A positive control (an individual with a *C9ORF72* repeat expansion) was run to ensure successful execution of the protocol.

## 4.3 Results

### 4.3.1 Ancestral background

To confirm that all Rapsani Greek individuals are of homogenous European descent, we ran a MDS analysis (Figure 57). As all samples from this cohort cluster uniformly, we can confirm their European ancestral origins.



**Figure 57: Multidimensional Scaling of Greek Rapsani village members.**

All individuals cluster uniformly with those of European descent. Key is located in the bottom right hand corner.

In addition to multi-dimensional scaling, we also measured the inbreeding coefficients (f) of the Rapsani family members to determine the expected likelihood of genetic effects due to inbreeding using pedigree relationships.

FID	IID	O(HOM)	E(HOM)	N(NM)	F
RGPD_78112	GPD_78112	947929	965700	1382757	-0.04255
GPD_78158	GPD_78158	945759	965200	1382064	-0.04661
GPD_78160	GPD_78160	946840	966700	1384347	-0.04764
GPD_78227	GPD_78227	947958	966700	1384307	-0.04491
GPD_78231	GPD_78231	950394	966600	1384206	-0.0389
GPD_78232	GPD_78232	948254	966500	1383955	-0.04365
GPD_78238	GPD_78238	944044	964500	1381151	-0.04921
R1F15	R1F15	948991	965900	1383098	-0.04046
R1F3	R1F3	950535	966400	1383839	-0.03798
R1F9	R1F9	943887	964200	1380593	-0.04871
R2F18	R2F18	948838	960000	1374502	-0.02691
R2F1	R2F1	941649	962300	1377873	-0.04971
R2F3	R2F3	950499	965300	1382273	-0.03551
R2F4	R2F4	947680	965900	1383172	-0.04378
R2F7	R2F7	950622	965500	1382505	-0.03563
R2F8	R2F8	947684	964600	1381250	-0.04065
R3F1	R3F1	944528	959600	1374003	-0.0364
R3F3	R3F3	945796	963900	1380233	-0.04361
R4F1	R4F1	951659	965800	1382991	-0.03385
RCON8	RCON8	946761	962400	1377968	-0.03753

**Table 31: Observed and expected rates of homozygosity and inbreeding coefficients of Rapsani villagers.**

The inbreeding coefficient (F) reflects the approximate percentage of homozygous alleles in an individual's genome. The larger the F value, the more biologically related an individual's parents are and the higher the probability of increased homozygosity. FID = Family ID. IID = Individual ID. O(HOM) = Observed number of homozygotes. E(HOM) = Expected number of homozygotes. N(NM) = Number of non-missing genotypes.

### 4.3.2 Genotyping

The quality control results of the genotyping data are listed below in Table 32.

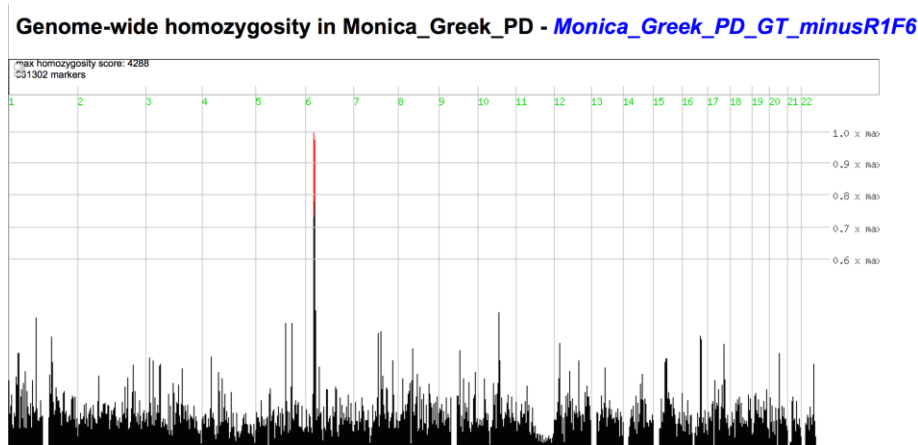
With the exception of sample R1F6, all samples generated high quality data and passed standard quality control measures.

Row	DNA_ID	#No_Calls	#Calls	Call_Rate	A/A_Freq	A/B_Freq	B/B_Freq	Minor_Freq	50%_GC_Score	10%_GC_Score	Pass/Fail
1	R1F15	6293	2561552	0.9975	0.3487	0.1697	0.4816	0.4335	0.7786	0.4169	P
2	R1F6	929494	1638351	0.638	0.3607	0.4078	0.2315	0.4354	0.6927	0.3145	F
3	R1F3	4834	2563011	0.9981	0.3485	0.1694	0.4821	0.4332	0.7788	0.4177	P
4	R1F9	10044	2557801	0.9961	0.3476	0.1711	0.4813	0.4332	0.7773	0.4156	P
5	R3F3	9859	2557986	0.9962	0.3483	0.1702	0.4815	0.4334	0.777	0.4151	P
6	R3F1	20670	2547175	0.992	0.3476	0.1729	0.4795	0.434	0.7752	0.4114	P
7	R2F3	8011	2559834	0.9969	0.3468	0.1728	0.4804	0.4332	0.7785	0.4168	P
8	R2F7	6724	2561121	0.9974	0.3486	0.169	0.4825	0.4331	0.7784	0.4169	P
9	R2F4	6490	2561355	0.9975	0.3461	0.1741	0.4798	0.4332	0.7788	0.4173	P
10	R4F1	7052	2560793	0.9973	0.3468	0.1726	0.4806	0.4331	0.7787	0.4171	P
11	RCON8	15593	2552252	0.9939	0.3477	0.173	0.4792	0.4343	0.7769	0.4135	P
12	R2F8	9229	2558616	0.9964	0.3464	0.1738	0.4799	0.4333	0.7779	0.416	P
13	R2F1	14547	2553298	0.9943	0.3463	0.1751	0.4785	0.4339	0.776	0.4132	P
14	GPD_78112	6924	2560921	0.9973	0.3458	0.174	0.4801	0.4328	0.7786	0.4174	P
15	R2F18	20198	2547647	0.9921	0.3503	0.1675	0.4822	0.434	0.7757	0.4133	P
16	GPD_78158	8512	2559333	0.9967	0.3457	0.1746	0.4797	0.433	0.7786	0.4169	P
17	GPD_78160	4852	2562993	0.9981	0.3453	0.175	0.4798	0.4328	0.7793	0.4184	P
18	GPD_78227	4221	2563624	0.9984	0.3475	0.1705	0.482	0.4327	0.7791	0.4181	P
19	GPD_78231	5120	2562725	0.998	0.3461	0.1735	0.4804	0.4329	0.7792	0.4182	P
20	GPD_78232	5594	2562251	0.9978	0.346	0.1742	0.4798	0.4331	0.779	0.4179	P
21	GPD_78238	8872	2558973	0.9965	0.3478	0.1712	0.481	0.4334	0.7769	0.4152	P

**Table 32: Quality control of Greek PD Rapsani cohort genotyping.**

Genotyping performed using the Illumina OmniExp-12,v1.0 DNA Analysis BeadChip. Highlighted in red R1F6 failed due to the very low call rate (63.8%), while all other individuals had call rates >99%

After removing R1F6 from the genotyping cohort, we looked for any shared regions of homozygosity. By assigning respective case and control status to our uploaded genotypes, a Manhattan plot of increased regions of homozygosity was generated (Figure 58).

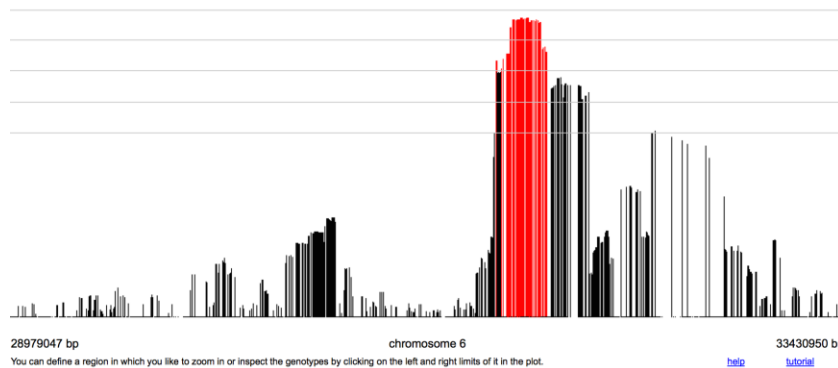


**Figure 58: Greek Rapsani cohort genotyping data viewed in homozygosity mapper**

Chromosomes are depicted numerically on the x-axis and regions of homozygosity are illustrated by peaks along the y-axis, with taller peaks reflecting genomic areas of increased homozygosity.

Initially the single peak on chromosome 6 appeared promising, suggesting a shared region of homozygosity based on common variation. Further analysis of this region revealed a distinct peak located on chr6: 31588450 - 31846823 (Figure 59).

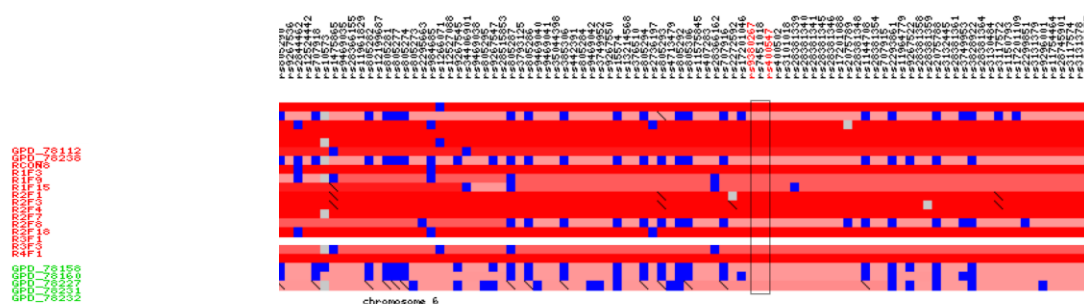
**Homozygosity on chromosome 6 in Monica\_Greek\_PD - *Monica\_Greek\_PD\_GT\_minusR1F6***



**Figure 59: Greek Rapsani cohort genotyping data viewed in homozygosity mapper.**

Zoomed-in view of distinct peak on chromosome 6.

A thorough analysis of each peak, including all of the genes and intergenic regions, revealed the absence of any significant findings. First and foremost, there were no homozygous stretches which were present in cases but absent in controls (or vice versa). However, given that some of the controls are relatively young and may develop PD later in life, we relaxed the criteria in our analyses and investigated any homozygous stretches in these regions, present in all cases and allowing presence of the identical region in controls. An example of a run of homozygosity among all samples is demonstrated in the image below (Figure 60).



**Figure 60: An example of a region of homozygosity along chromosome 6 for all Rapsani individuals.**

The top row represents genotyped SNPs along chromosome 6. The side panel on the left-hand side represents all individuals genotyped, in parallel to each colored row. The first 16 rows represent all cases and the bottom 7 rows denote all unaffected individuals, with the white horizontal line separating case/control cohorts, respectively. Regions in red reflect purely homozygous regions, while those in blue represent heterozygous genotypes. Downward diagonal black slashes denote that an individual is homozygous for the minor allele.

On chromosome 6, several regions of homozygosity were located within either intergenic regions or pseudo-genes and the remaining 25 resided within different domains of protein-coding genes. Notably, however, no other regions in the genome demonstrated runs of homozygosity. The complete list of 31 genes within this domain was obtained using the gene distiller function. Among the 25 genes demonstrating runs of homozygosity within protein-coding genes, 6 (highlighted in yellow) appeared to be in exonic regions (Table 33).



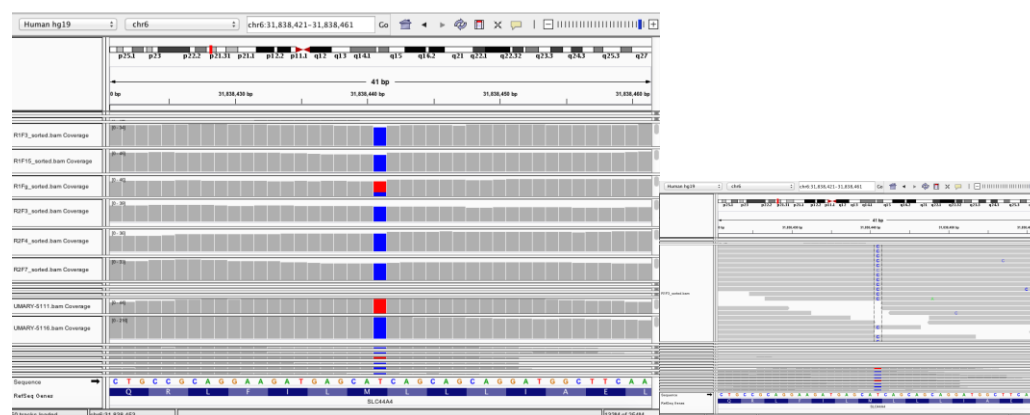
Gene	Name	Type	KEGG pathways, associations
<b>PRRC2A</b>	<b>proline-rich coiled-coil 2A</b>	protein-coding	<b>HLA-B</b>
BAG6	BCL2-associated anthanogene 6	protein-coding	degradation of mislocalized proteins, ubiquitin pathway
APOM	apolipoprotein M	protein-coding	triglycerides, lipoproteins
<b>C6orf147</b>	<b>chr6 open reading frame 47</b>	protein-coding	<b>unknown</b>
GPANK1	G patch domain and ankryin repeats 1	protein-coding	HLA-B
CSNK2B	Casein kinase 2, beta polypeptide	protein-coding	Wnt signaling, Adherens and Tight junctions
LY6G5B	lymphocyte antigen 6 complex, locus G5B	protein-coding	MHC class III
LY6G5C	lymphocyte antigen 6 complex, locus G5C	protein-coding	MHC class III
ABHD16A	abhydrolase domain containing 16A	protein-coding	HLA-B
LY6G6F	lymphocyte antigen 6 complex, locus G6F	protein-coding	immunoglobulin superfamily
LY6G6D	lymphocyte antigen 6 complex, locus G6D	protein-coding	MHC class III
LY6G6C	lymphocyte antigen 6 complex, locus G6C	protein-coding	MHC class III
C6orf25	chr6 open reading frame 25	protein-coding	platelet and immunoglobulins
DDAH2	dimethylarginine dimethylaminohydrolase 2	protein-coding	Nitric Oxide Synthase (NOS) reactions
CLIC1	chloride intracellular channel 1	protein-coding	homeostasis, metabolism, hematopoietic system
MSH5	mutS homolog 5	protein-coding	DNA mismatch repair, meiotic recombination
SAPCD1	suppressor APC domain containing 1	protein-coding	unknown
VWA7	von Willebrand factor A domain containing 7	protein-coding	HLA class III
<b>VARS</b>	<b>valyl-tRNA synthetase</b>	protein-coding	<b>Val, Leu, Ile biosynthesis, Aminoacyl-tRNA biosynthesis</b>
LSM2	LSM2 homolog, U6 snRNA and mRNA degradation associated	protein-coding	RNA degradation, spliceosome
<b>HSPA1L</b>	<b>heat shock 70kDA protein 1-like</b>	protein-coding	<b>spliceosome, MAPK signaling pathway, Endocytosis, Antigen processing</b>
<b>HSPA1A</b>	<b>heat shock 70kDA protein 1A</b>	protein-coding	<b>same as above plus prion diseases</b>
C6orf48	chr6 open reading frame 48	protein-coding	MHC class III
NEU1	sialidase 1 (lysosomal sialidase)	protein-coding	sphingolipid metabolism, lysosome, other glycan degradation
<b>SLC44A4</b>	<b>solute carrier family 44, member 4</b>	protein-coding	<b>transmembrane transport, metabolism, synthesis of small molecules</b>

**Table 33: Homozygosity mapper gene distiller output of Greek Rapsani cohort.**

Includes only protein-coding genes (25) on chromosome 6 that illustrated distinct runs of homozygosity within the chromosome 6 peak. Those highlighted represent regions of homozygosity within coding regions of their respective gene.

This region of homozygosity lies on chromosome 6, across the MHC region, consequently several genes within this run of homozygosity are highly associated with immune functions, including the Major Histocompatibility complex (MHC) III and autoimmune disorders like Systemic lupus erythematosus (SLE). This region is known to show population specific selection and notably all sampled individuals, including all controls, were identically homozygous across this region; thus we felt that this region was unlikely to contain the disease-causing variant. However, we investigated this region carefully using the available sequence data.

After scrutinizing all exons of the genes within the homozygous chromosome 6 region in all individuals, all variants of initial interest appeared to be artifacts and/or to be present in a series of control samples run as part of a separate project, thus no variant that fit our criteria for likely disease causing was identified (Figure 61).



**Figure 61: Example of artifact of novel SLC44A4 variant based on BAMs**

From the image on the left, all affected individuals who were both whole exome and whole genome sequenced appear to be homozygous or heterozygous for a novel variant in SLC44A4. However, as UMARY laboratory samples were also used as negative controls, this “call” is clearly an artifact. The image on the right reveals the excellent paired end read coverage and high confidence calls in R1F3 based on the Phred score; nonetheless, this represents a technical artifact inherent to both WES and WGS technologies.

In addition to analyzing data using homozygosity mapper we also used PLINK to run two similar tests: first, aimed at detecting runs of homozygosity and second looking for shared segments, irrespective of whether the segments were in homozygous regions or not. While the former yields information analogous to that obtained above using homozygosity mapper, the latter can identify shared segments of homozygous or heterozygous genotypes based on common genotypic variation. By looking at each chromosome individually with phased genotypes at -1Mb windows, we were unable to identify any shared regions between all individuals (Table 34).

ID I	ID II	CHROM	Segment start (bp)	Segment end (bp)	Segment start (SNP)	Segment end (SNP)	Total SNPs in segment	Genetic length of segment	Units for genetic length	Mismatching SNPs in segment	*	**
GPD_78160	R1F9	1	33170413	212298845	rs9425961	rs9804026	512	179.13 MB			1	0
R2F3	R2F4	1	100149952	179852074	rs12136177	rs1281378	1280	79.7 MB			5	0
R2F3	R2F4	1	66534276	23983351	rs1392818	rs16836647	1024	173.3 MB			1	0
R2F18	R2F8	1	66534276	162874838	rs1392818	rs16846617	1152	96.34 MB			6	0
R2F3	R2F4	1	114380395	117503940	exm85438	exm87515	256	3.12 MB			0	0
GPD_78227	GPD_78232	10	21239877	92442274	rs1856762	rs1892105	128	71.2 MB			0	0
R2F18	R2F8	10	7286305	113230705	kgp21985014	kgp21677484	1152	105.94 MB			6	0
R2F3	R2F4	10	7286305	113230705	kgp21985014	kgp21677484	1152	105.94 MB			2	0
R2F3	R2F4	11	25117751	110375034	rs12805702	rs1553734	640	85.26 MB			1	0
R2F18	R2F8	11	28536507	33689529	kgp12659553	kgp12705874	256	5.15 MB			1	0
R2F3	R2F4	11	461538	99419823	kgp12612318	kgp12853609	1280	98.96 MB			5	0
R2F3	R2F4	11	14240902	120313682	kgp12946006	kgp22774772	1024	106.07 MB			1	0
GPD_78158	R2F7	11	36595600	102826412	exm900950	exm951629	6144	66.23 MB			8	0
R2F3	R2F4	12	15768493	95591586	rs7294628	rs7956997	1280	79.82 MB			1	0
R3F1	R3F3	12	47546346	65889778	kgp7144937	kgp7946078	2560	18.34 MB			9	0
R2F3	R2F4	13	64906263	106001151	rs9540053	rs9558536	640	41.09 MB			1	0
R2F3	R2F4	13	37110899	78215357	kgp11649952	kgp16591508	1280	41.1 MB			1	0
GPD_78160	R1F9	13	19706775	84716696	kgp16624439	kgp16669550	896	65.01 MB			3	0
R2F3	R2F4	13	23201300	4630047	kgp16833058	kgp16894637	1280	23.63 MB			5	0
R2F3	R2F4	14	24402725	67682370	rs8011460	rs8012740	384	43.28 MB			2	0
R2F3	R2F4	14	42759494	96293279	kgp19543721	kgp19638746	2048	53.53 MB			6	0
R2F3	R2F4	15	46912093	86767418	kgp10095813	kgp10675756	1280	39.86 MB			5	0
R2F3	R2F4	16	55366317	77189406	kgp16455644	kgp22806065	640	21.82 MB			2	0
GPD_78160	R1F9	16	80231788	83515062	kgp7147119	kgp11573477	640	3.28 MB			1	0
GPD_78160	R1F9	16	48899847	81531606	kgp16221210	kgp16252241	640	32.63 MB			1	0
GPD_78160	R1F9	16	5612223	24487674	kgp16419264	kgp16443401	512	18.88 MB			0	0
R2F18	R2F8	16	73265603	88497400	kgp8662552	kgp10003256	256	15.23 MB			0	0
GPD_78160	R1F9	17	7387210	77275868	kgp13970330	kgp13998036	640	3.89 MB			1	0
R3F1	R3F3	17	50365439	77275868	kgp13992393	kgp13998036	128	26.91 MB			0	0
R2F3	R2F4	18	53639716	62045482	kgp10812446	kgp15935985	640	8.41 MB			1	0
GPD_78160	R1F9	18	38560259	44494375	kgp15959888	kgp15984213	512	5.93 MB			1	0
GPD_78160	R1F9	18	56131349	73494564	kgp16102189	kgp2682608	896	17.36 MB			4	0
R2F18	R2F8	18	3636125	11671835	kgp1121123	kgp11266443	128	8.04 MB			0	0
R2F3	R2F4	18	49629625	72004071	kgp11023402	kgp11611417	1280	22.37 MB			5	0
GPD_78160	R1F9	19	35820889	51378273	kgp11355004	kgp11801100	640	15.56 MB			1	0
R2F3	R2F4	19	32680545	52839356	kgp21408523	kgp21484566	1280	20.16 MB			1	0
R2F3	R2F4	5	59890278	87038565	rs4326096	rs710384	256	27.15 MB			0	0
GPD_78160	R1F9	5	23139973	72023208	rs10038745	rs10050720	256	48.88 MB			0	0
GPD_78160	R1F9	5	95601922	161212950	rs12109041	rs1658074	640	65.61 MB			1	0
R2F3	R2F4	5	71295157	153000753	rs13187253	rs13358432	128	81.71 MB			0	0
R3F1	R3F3	5	26819303	67238423	rs2248868	rs2372189	256	40.42 MB			1	0
R2F3	R2F4	5	82685260	106130200	rs2441010	rs296277	1408	23.44 MB			2	0
R3F1	R3F3	5	3765162	76543867	kgp22579137	kgp22669471	2304	72.78 MB			7	0
R2F3	R2F4	5	172179222	178096520	kgp3151455	kgp3230592	384	5.92 MB			2	0
R2F3	R2F4	6	27551455	139757582	rs9295737	rs10085329	1280	112.21 MB			1	0
R2F18	R2F8	6	112536990	160196343	rs2213840	rs25683	1152	47.66 MB			6	0
R2F3	R2F4	6	112536990	160196343	rs2213840	rs25683	1152	47.66 MB			2	0
R3F1	R3F3	6	2859121	148556854	kgp17034652	kgp17130411	2560	145.7 MB			9	0
R2F18	R2F8	6	151914326	167570961	exm587195	exm594365	1024	15.66 MB			3	0
R2F3	R2F4	7	37604528	148480990	rs10257469	rs10271133	384	110.88 MB			2	0
GPD_78160	R1F9	7	17911038	107992582	rs10282707	rs10487851	256	90.08 MB			0	0
R2F3	R2F4	8	84114931	94019491	rs15685328	rs16915833	640	86.62 MB			1	0
GPD_78160	R1F9	8	2412280	89030490	rs17670319	rs1778917	640	86.62 MB			1	0
GPD_78227	GPD_78232	8	2734916	141550634	rs2604146	rs2944765	768	138.83 MB			3	0
R2F3	R2F4	8	37480642	75254086	rs6468423	rs7004754	1408	37.77 MB			2	0
R3F1	R3F3	8	20685010	127772865	kgp4404589	kgp4935987	2048	107.09 MB			5	0
GPD_78160	R1F9	9	14498764	33420271	kgp10051302	kgp18371735	896	18.92 MB			4	0
R2F18	R2F8	9	81516107	132785679	kgp18247351	kgp18254288	128	51.27 MB			0	0
R2F18	R2F8	9	121639006	135679627	kgp18323400	kgp18387514	1152	14.04 MB			6	0
R2F3	R2F4	9	121639006	135679627	kgp18323400	kgp18387514	1152	14.04 MB			2	0
R2F3	R2F4	X	35245289	42862609	kgp22784659	kgp22798289	640	7.62 MB			1	0
GPD_78227	GPD_78232	X	46966776	86322628	kgp22730192	kgp22730786	128	39.36 MB			0	0
R2F3	R2F4	X	24157207	71864857	kgp22743558	kgp22749785	1280	47.71 MB			1	0
GPD_78160	R1F9	X	92090301	150748477	kgp22752720	kgp22756802	896	58.66 MB			3	0
R2F18	R2F8	X	26038762	83001032	kgp22762317	kgp22765904	128	57.86 MB			0	0
R2F3	R2F4	X	2255142	123709758	kgp22780827	kgp22789007	1024	121.45 MB			1	0
R2F3	R2F4	18	53639716	62045482	kgp10812446	kgp15935985	640	8.41 MB			1	0
GPD_78160	R1F9	18	38560259	44494375	kgp15959888	kgp15984213	512	5.93 MB			1	0
GPD_78160	R1F9	18	56131349	73494564	kgp16102189	kgp2682608	896	17.36 MB			4	0
R2F18	R2F8	18	3636125	11671835	kgp1121123	kgp11266443	128	8.04 MB			0	0
R2F3	R2F4	18	49629625	72004071	kgp11023402	kgp11611417	1280	22.37 MB			5	0
GPD_78160	R1F9	19	35820889	51378273	kgp11355004	kgp11801100	640	15.56 MB			1	0
R2F3	R2F4	19	32680545	52839356	kgp21408523	kgp21484566	1280	20.16 MB			1	0
GPD_78160	R1F9	19	3344483	19886812	kgp21527058	kgp22750074	896	16.54 MB			4	0
R2F18	R2F8	19	36292754	58791636	kgp4230340	kgp4337778	128	22.5 MB			0	0
R2F3	R2F4	2	47952231	235480729	rs797689	rs9287578	640	187.53 MB			1	0
GPD_78160	R1F9	2	11359120	147342341	rs9808232	kgp12432465	512	135.98 MB			1	0
GPD_78158	R2F7	2	27441640	177296916	kgp14449358	kgp14689212	6144	149.86 MB			8	0
R2F3	R2F4	20	23777887	55809932	kgp19324252	kgp19345624	512	32.11 MB			0	0
R2F18	R2F8	20	14518540	31679301	kgp22802358	kgp2676221	128	17.16 MB			0	0
GPD_78160	R1F9	20	34595436	52687027	kgp2273226	kgp50695	640	18.09 MB			1	0
R2F18	R2F8	20	33226491	43610458	kgp7423007	kgp19192952	512	10.38 MB			2	0
R2F3	R2F4	20	14513492	36581445	kgp10381737	kgp11128255	1280	22.07 MB			1	0
R2F18	R2F8	20	5082564	35811176	kgp19257794	kgp19263893	128	30.73 MB			0	0
GPD_78160	R1F9	21	32768399	43797515	kgp3631732	kgp4340473	640	11.03 MB			1	0
GPD_78160	R1F9	22	17600375	25010855	exm1583099	exm1594803	896	7.41 MB			4	0
R2F3	R2F4	3	3968208	60589986	rs10865884	rs1158425	512	56.62 MB			1	0
R2F3	R2F4	3	173310353	182650847	rs17178976	rs2089612	1280	9.34 MB			1	0
GPD_78160	R1F9	3	10285888	189901117	rs242724	rs3107688	896	179.62 MB			4	0
R2F3	R2F4	3	103097277	187935634	rs6441729	rs6785028	1280	84.84 MB			5	0
R3F1	R3F3	3	64512088	102779892	kgp17683247	kgp17779266	2560	38.27 MB			9	0
R2F18	R2F8	3	38545177	46496854	exm300976	exm307868	1024	7.95 MB			3	0
R2F3	R2F4	4	89065353	164884738	rs2622627	rs3967462	1280	75.82 MB			1	0
GPD_78160	R1F9	4	101579972	141870648	rs4586953	rs4956494	896	40.29 MB			4	0
R2F18	R2F8	4	125017034	138397480	rs12647636	rs10032306	1152	13.38 MB			6	0
R2F3	R2F4	4	125017034	138397480	rs12647636	rs10032306	1152	13.38 MB			2	0

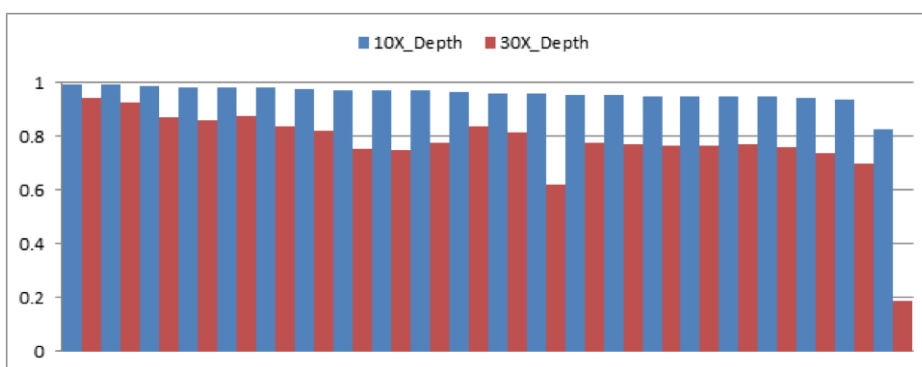
Table 34: Results of “Shared Segment analysis” using Plink and GERMLINE.

Kgp = non-polymorphic CNV markers. \* 1 if Individual 1 is homozygous in match; 0 otherwise. \*\*1 if Individual 2 is homozygous in match; 0 otherwise

### 4.3.3 Whole exome sequencing

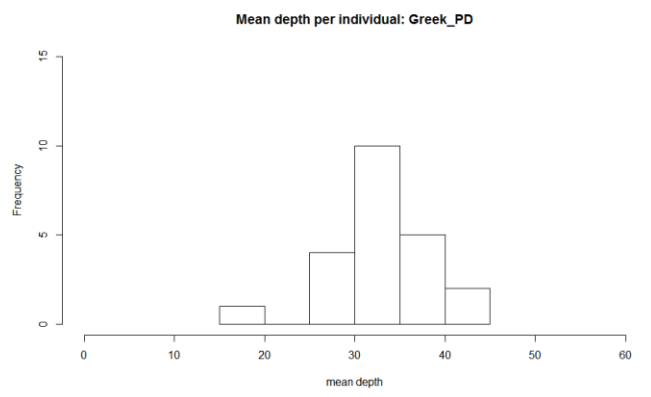
#### 4.3.3.1 Quality control filtering

First the quality of exome sequencing data was assessed by looking at key parameters, including depth, coverage and PCR duplicate rates. The samples that did not meet the following criteria were eliminated: >90% 10X depth, >70% 30X depth and a PCR duplicate rate <14% (Figure 62).



**Figure 62: Whole exome sequencing 10X and 30X depth of Greek Rapsani village samples.**

All data was obtained using the Illumina Nextera Protocol. Each Rapsani individual is represented by a single blue and a single red line on the x-axis. Each line corresponds to the 10x (blue) and 30x (red) depths, with coverage values on the y-axis ranging from 0-1, with 1 representing 100% coverage.



**Figure 63: Whole exome sequencing: mean depth per individual in 22 Greek Rapsani PD samples**

Mean depth was approximately 33X per individual.

#### 4.3.3.2 Examining known PD genes

According to the variant filtering pipeline in the methods section (Figure 54), the first step in data analysis after quality control measures was to identify any variants in genes associated with PD. Among all 23 individuals, a total of 110 variants were identified in 43 PD associated genes. After a thorough search on Human Molecular Genomic Database (HMGD), 9 of these variants (in 4 genes) were considered PD associated polymorphisms but none were considered disease causal and thus we were unable to explain the disease in this population (Table 35). Notably, this method would be insensitive to structural mutations such as those found at the *SNCA*, *PARK2*, *PINK1*, or *DJ1* loci, however these regions did not show up in the homozygosity or shared segment analyses and thus are an unlikely cause of disease in this population.

MAF	Cases (HOM)	Cases (HET)	CNTRLs (HOM)	CNTRLs (HET)	Gene and corresponding transcript ID
0.092	0	3	0	0	PINK1:NM_032409:exon5:c.G1018A:p.A340T,
0.027	0	1	0	0	HTRA2:NM_145074:exon1:c.G421T:p.A141S,HTRA2:NM_013247:exon1:c.G421T:p.A141S,
0.026	0	0	0	2	PARK2:NM_004562:exon11:c.G1180A:p.D394N,PARK2:NM_013987:exon10:c.G1096A:p.D366N,PARK2:NM_013988:exon8:c.G733A:p.D245N,
0.165	0	7	0	2	PARK2:NM_004562:exon10:c.G1138C:p.V380L,PARK2:NM_013987:exon9:c.G1054C:p.V352L,PARK2:NM_013988:exon7:c.G691C:p.V231L,
0.068	0	1	0	0	PARK2:NM_004562:exon4:c.G500A:p.S167N,PARK2:NM_013987:exon4:c.G500A:p.S167N,
0.086	0	2	0	1	LRRK2:NM_198578:exon14:c.C1653G:p.N551K,
0.084	0	2	0	1	LRRK2:NM_198578:exon30:c.G4193A:p.R1398H,
0.298	1	7	1	2	LRRK2:NM_198578:exon34:c.T4939A:p.S1647T,
4.1E-07	0	1	0	0	VPS35:NM_018206:exon17:c.C2320A:p.L774M,

**Table 35: All PD-associated polymorphisms identified in 23 members of Greek Rapsani cohort.**

All PD-associated polymorphisms identified in 23 members of Greek Rapsani cohort. HOM = homozygous. HET= heterozygous. CNTRLs= controls. Numbers refer to how many individuals carry each variant.

#### 4.3.3.3 Population database filtering

After determining that all known PD mutations were absent among the entire Greek Rapsani cohort, we proceeded with variant filtering according to the pipeline discussed in the methods section (Figure 54). This entailed extensive population database

filtering based on those listed in Figure 29 (1000 genomes, dbsnp, ESP6500). With 4184 variants remaining, we used two distinct analytic approaches, the former being more stringent and the latter significantly more lenient.

#### 4.3.3.3.1 *Stringent Filtering Pipeline:*

Subsequent to population database filtering, we performed stringent prediction site database filtering, resulting in 414 variants. Among these 414 variants, 26 appeared to be shared by all affected individuals according to the gVCF. After visualizing the BAMs for all 26 variants in each Greek Rapsani sample on IGV, we excluded 23 variants as likely sequence and alignment artifacts and selected 3 variants that appeared promising in the following genes: *CEP290*, *KMT2C* and *CTSF*. The variants were located in genes exhibiting neuronal expression in the brain according to GeneCards (<http://www.genecards.org/>) and interacted with proteins related to PD or PD-associated mechanisms (i.e. autophagy) using STRING and KEGG. Sanger sequencing revealed that variants in *CEP290* and *KMT2C* were artifacts and only 3 affected individuals (out of 16) were heterozygous for the variant in *CTSF*. As this filtering process may have been too harsh, perhaps removing plausible candidate genes, we moved towards a more lenient approach.

#### 4.3.3.3.2 *Liberal Filtering Pipeline*

We started again with 4184 variants following population database filtering (Figure 54). Next, we filtered by MAF, keeping only variants with a MAF<1%. This resulted in 333 variants in total, with 24 homozygous variants and 309 heterozygous

variants remaining. However, none of these 333 variants were shared by all affected individuals. Acknowledging that each family may carry a unique pathological variant, we pursued individual family analyses for the three larger Rapsani families (I-III). From the 333 variants in the previous step, all variants with a MAF >.1% were filtered out, resulting in a total of 200 candidate variants. Secondly, as Rapsani families I and II consist of 4 and 7 affected members, respectively, we removed all variants that were not shared by at least 3 affected members of each family. Since Rapsani family 3 only consists of 2 members, we required all variants to be shared by both members. This resulted in 51 variants remaining for individual familial analyses. After visualizing all 51 variants on IGV, 25 variants were selected for Sanger sequencing based upon BAM appearance and at least some neuronal expression in the brain. A list of these variants and their validation outcomes are listed below.

Chr	POS	ID	REF	ALT	Gene	Effect	MAF	Rapsani families	# Individuals heterozygous
chr16	25251790	.	G	C	ZKSCAN2	NS SNV			artifact
chr4	266147	rs202099832	T	C	ZNF732	NS SNV	0.07%		artifact
chr9	14859241	.	C	T	FREM1	NS SNV			artifact
chr11	125322281	.	C	T	FEZ1	NS SNV		II	2
chr11	95826086	rs191030213	G	A	MAML2	NS SNV	0.03%	II	1
chr2	236626225	rs201400274	G	A	AGAP1	NS SNV	0.09%	I, II	4
chr12	58120546	.	T	C	AGAP2	NS SNV			artifact
chr15	75042189	.	A	G	CYP1A2	NS SNV		I, II	3
chr12	22199386	.	T	C	CMAS	NS SNV			artifact
chr17	1386937	rs146471734	C	T	MYO1C	NS SNV	0.0053%	I, II, III	5
chr14	101005265	.	C	T	BEGAIN	NS SNV		I, II, V	4
chr19	54409660	.	C	A	PRKCG	NS SNV			artifact
chr12	57351191	rs199814447	C	T	RDH16	NS SNV	0.0066%		artifact
chr4	153896551	rs147950044	C	T	FHDC1	NS SNV	0.0025%	III	2
chr12	118693338	rs78662524	G	T	TAOK3	NS SNV	0.07%	III	
chr7	6217468	.	G	C	CYTH3	NS SNV		III	
chr6	158517100	.	G	C	SYNJ2	NS SNV		III	
chr16	15703534	rs200431093	A	G	KIAA0430	NS SNV	0.05%	III	
chr16	30774746	.	A	T	RNF40	NS SNV		III	
chr19	44738864	.	A	T	ZNF227	NS SNV		III	
chr15	91454462	rs200259502	C	T	MAN2A2	NS SNV	0.02%	III	
chr16	18865003	rs184038326	C	T	SMG1	NS SNV	0.07%	III	
chr3	108366806	.	G	C	DZIP3	NS SNV		III	
chr5	148427540	.	G	T	SH3TC2	NS SNV		III	
chr17	74037057	.	G	C	SRP68	NS SNV		III	

**Table 36: WES sanger sequencing results for individual Rapsani families**

Those highlighted in yellow were frozen as clean PCR plates and put on hold, as we decided that we would pursue whole genome sequencing and re-analyze the data.

The Sanger sequencing results revealed several artifacts but also confirmed multiple variants. While none of the 7 unaffected individuals carry any of these variants, not all affected individuals (even within a single family-aside from Rapsani III, which only has 2 members) possess validated variants. Thus, we questioned if the coding regions were inadequately covered, hindering our ability to identify a novel mutation. Further, we also entertained the possibility of SVs or CNVs as a cause of disease, resorting to WGS as a logical next step in analysis.

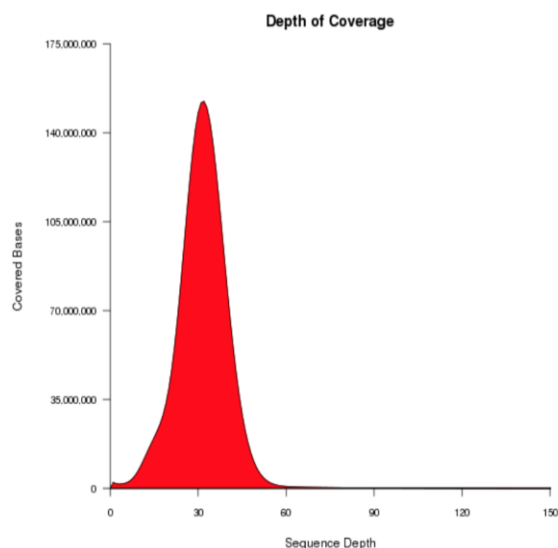
#### **4.3.4 Whole genome sequencing**

Given the high cost of WGS, we outsourced 7 of the 16 affected samples to MacroGen for WGS. This included those from the two larger pedigrees: 4 individuals from Rapsani I and 3 individuals from Rapsani II.

##### **4.3.4.1 Quality Control Filtering**

In line with exome data raw data processing, the depth and coverage were key parameters to assess sample quality control. As WGS target values are significantly higher than those for WES, the 10X coverage ( $>98.6$ ) was excellent, while the 30X ( $>61$ ) coverage was slightly lower than ideal, as we usually strive for  $>70$  in WES. However, the overall average depth of 30X was significantly greater than data obtained from WES, as expected (Figure 64).





% Coverage	%>1X	%>5X	%>10X	%>15X	%>20X	%>30X
Value	99.5	99.2	98.6	96.3	91.8	61.0

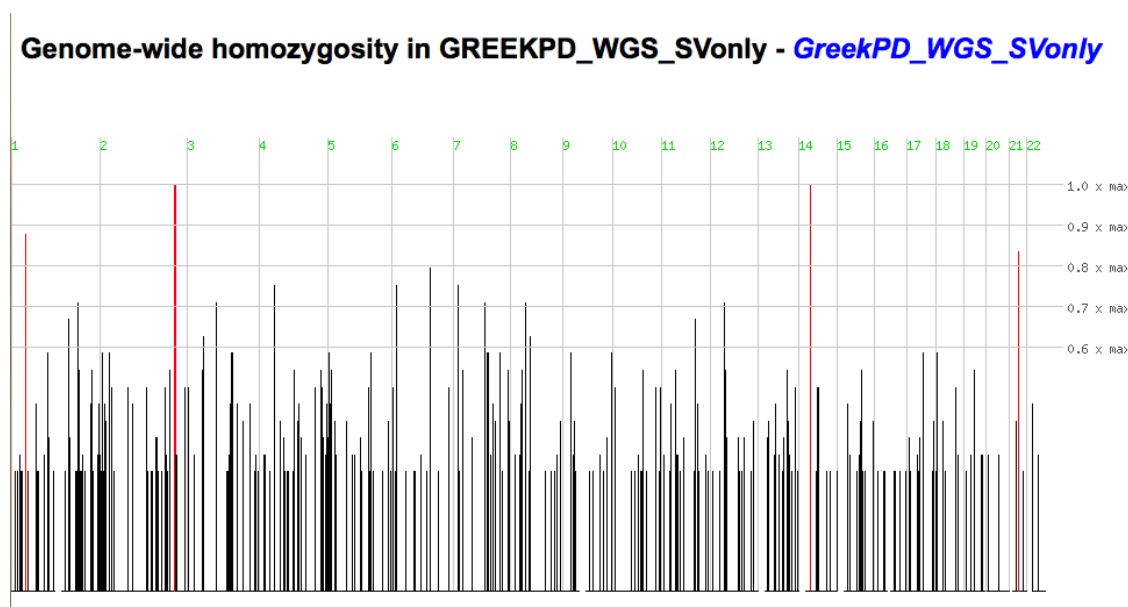
**Figure 64: Greek Rapsani PD whole genome sequencing quality control results.**

% Coverage reflects the percentage of bases in non-N reference regions with specific depth of coverage or greater.

#### 4.3.4.2 Homozygosity Mapper Analyses

##### 4.3.4.2.1 Structural variant homozygosity results

Before starting our filtering process with the SV, CNV and SNP-indel VCFs, it was important to upload and run the results on homozygosity mapper to identify runs of homozygosity in regions not covered by exome or genotyping data. We first looked at the SV results, which represented an entirely new data set, as SVs are not covered in WES. The results are illustrated below in Figure 65, revealing peaks on chromosomes 1, 2, 14, and 21.



**Figure 65: Greek Rapsani cohort WGS SV data viewed in homozygosity mapper**

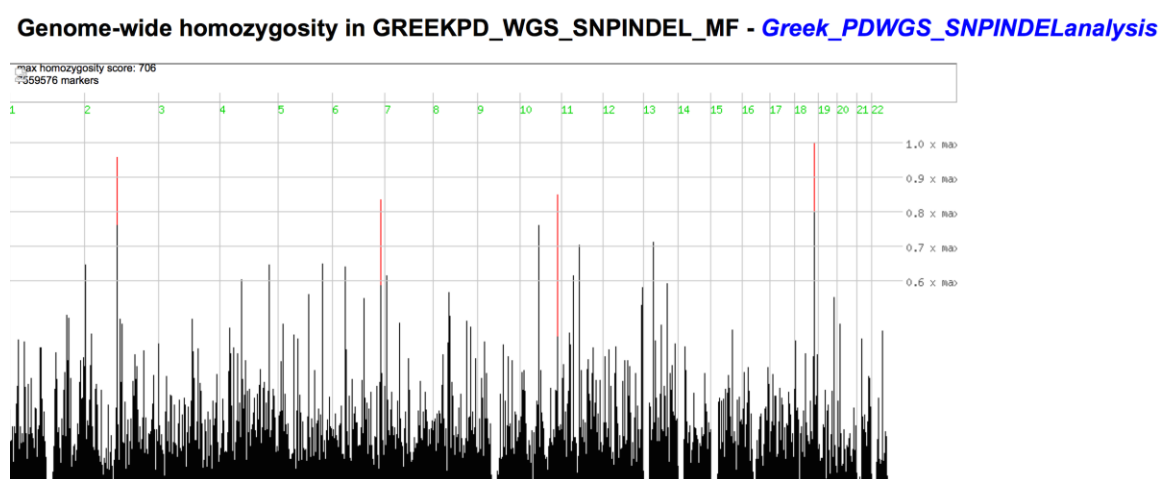
Chromosomes are depicted numerically on the x-axis and regions of homozygosity are illustrated by peaks along the y-axis, with taller peaks reflecting genomic areas of increased homozygosity.

Resonating with our previous analyses of genotyping data in homozygosity mapper, each peak was carefully analyzed to identify coding regions of protein-coding genes demonstrating runs of homozygosity among all affected members, as we did not obtain WGS data for any unaffected individuals. All novel or rare coding variants in runs of homozygosity were visualized on IGV; none of the variants, however, appeared real and thus we did not pursue validation.

#### 4.3.4.2.2 SNP indel homozygosity results

We also performed a homozygosity mapping analysis on the SNP and Indel gVCF. While this data includes exome results obtained by WES, it also includes coding regions that were not adequately captured (if at all). In addition, we could also identify homozygous regions within introns and UTRs. While coding regions are our primary

focus, this additional data can be beneficial toward identifying regions of linkage. As demonstrated below in Figure 66, 4 chromosomal peaks were visualized on chromosomes: 2, 6, 10 and 18.



**Figure 66: Greek Rapsani cohort WGS SNP indel data viewed in homozygosity mapper**

Chromosomes are depicted numerically on the x-axis and regions of homozygosity are illustrated by peaks along the y-axis, with taller peaks reflecting genomic areas of increased homozygosity. \*Note: The peak on chromosome 6 is in a different location than the homozygosity mapping results using genotyping data.

Each region was carefully evaluated for runs of homozygosity within coding regions of protein-coding genes; however, none of these regions could be verified by the WGS BAMs on IGV nor did any exhibit perfect segregation among cases alone. Furthermore, intronic and UTR regions were scrutinized for linkage and segregation, but also did not yield any findings.

#### **4.3.4.3 Whole genome sequencing variant filtering pipeline analysis**

As the WGS dataset is significantly larger than that obtained from WES, we defined specific strategies to analyze the following three data sets: CNVs, SVs and SNP-Indel VCFs.

#### 4.3.4.3.1 Copy number variant data analysis

As illustrated in Figure 56, the first step of CNV analysis was to determine the total number of CNVs among 7 WGS samples, which comprised 854 losses and 1630 gains. Logically the next step was to identify CNV losses and gains shared by all 7 affected individuals. This reduced the list of variants to 44 CNV losses and 92 CNV gains. As this was still a substantial number to investigate, located within all different genomic regions, we looked towards the SV data to identify any chromosomal breakpoints.

#### 4.3.4.3.2 Structural variant data analysis

Referring back to Figure 56, a total of 22,342 structural variants were obtained from WGS data among 7 affected Rapsani individuals. Among these, 650 variants were shared between all 7 samples: 96 heterozygous, 357 homozygous and 197 a mixture of both. Using strict filtering with population databases and prediction filters, 30 SVs remained. After scrutinizing BAMs of all 7 individuals, as well as UMARY controls on IGV, we pursued Sanger sequencing for 1 small SV in *DACHI*. This variant, however, failed to confirm.

#### 4.3.4.3.3 SNP Indel analysis

Among the 7 affected individuals, we started with almost 8 million variants (Figure 56). Using population databases (Figure 29), we filtered out several million variants, yielding a total of 790,519 variants. Analogous to SNV analysis with WES data, we took two analytical approaches, the former being more stringent and the latter significantly more liberal.

#### 4.3.4.3.3.1 SNP Indel stringent filtering approach

As we first wanted to identify a rare mutation shared by all 7 individuals from 2 Rapsani families, we filtered out all those not carried (in a heterozygous or homozygous state) by all samples. As we don't know the pattern of inheritance, it was important to include both types of alleles in our analysis. This left us with 30,779 shared variants. Next, using very harsh prediction filters including CADD, we were left with only 91 variants, 10 of which were located in protein-coding regions. Among these 10 variants, 3 (1 located in *SLAIN1*, 2 in *HYDINI*) appeared real in IGV, using several UMARY control WES BAMs as a standard of comparison. All 3 variants were followed up with sanger sequencing and determined to be artifacts.

#### 4.3.4.3.3.2 SNP-Indel liberal filtering approach

Cognizant that we only had 10 remaining protein-coding candidates in the previous analysis, we utilized mild prediction filters from the 790,519 variants obtained subsequent to population database filtering. This entailed the inclusion of all coding variants with a  $MAF < 1\%$  (including novel), resulting in coding 2337 variants, 1709 of which were non-synonymous. Next, we determined how many of these were shared among all 7 affected individuals, in either a heterozygous or homozygous state. However, we once again determined that all 7 individuals did not share a single variant among the 1709 non-synonymous variants, with a maximum only shared by 6 samples. While it was plausible that there was insufficient coverage for a sample in a particular region, we decided to investigate all non-synonymous variants present in at least 5 of the 7 individuals. Yielding only 23 variants, all of which were carefully visualized on IGV using UMARY samples as negative controls, we followed-up with sanger sequencing on

those that appeared real and exhibited at least some neuronal expression in the brain (according to GeneCards). A table of the 4 variants pursued with their results is listed below in Table 37. While some confirmed in several individuals, none validated in all 7 affected samples.

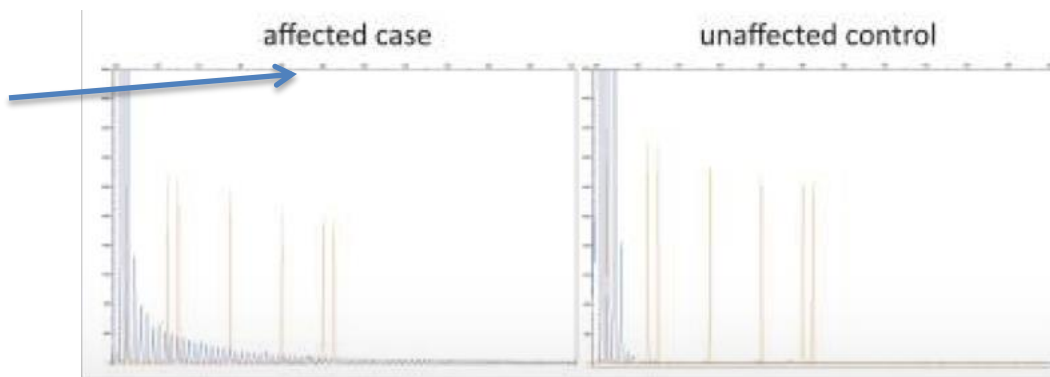
CHR	POSITION	ID	REF	ALT	GENE	EFFECT	Transcript	# Individuals with variant (out of 7)
chr9	70912543	-	A	T	<i>CBWD3</i>	Non-synonymous coding	NM_201453:exon12:Amic Acid:288	0
chr13	72440658	-	TGCCGCC	T	<i>DACHI</i>	Codon deletion	NM_080759:exon1:c.244_249del:p.82_83del, NM_004392:exon1:c.244_249del:p.82_83del	0
chr14	67940153	-	T	C	<i>TMEM229B</i>	Non-synonymous coding	NM_182526:exon3:c.A488G:p.H163R	1
chr14	68038891	-	C	G	<i>PLEKHH1</i>	Non-synonymous coding	NM_020715:exon11:c.C1625G:p.A542G	1

**Table 37: Sanger sequencing results of variants identified using a WGS liberal filtering approach**

Among the two variants that validated, neither reflects a worthwhile candidate to pursue in individual family analyses, given that only one sample out of 7 confirmed for each.

#### 4.3.5 *C9ORF72* screening

As at least one of the affected individuals in the Rapsani family cohort also presented with Motor Neuron disease (MND), we decided to screen for the intronic hexanucleotide repeat expansion in *C9ORF72*, which is a cause of familial ALS and FTD. An example of a positive and negative control is shown below in Figure 67.



**Figure 67: Example of positive and negative control for *C9ORF72* hexanucleotide repeat screening.**

Reproduced from (Renton et al 2011). The graphs depict capillary-based sequence traces of the repeat-primed PCR in an affected and unaffected sample. Orange lines reflect the size markers, and the vertical axis denotes the fluorescence intensity. In the affected individual, there is a classic “saw tooth tail” pattern that extends beyond the 300 bp marker with a 6 bp periodicity. This pattern is typical for affected individuals carrying the GGGGCC repeat expansion.

We analyzed our results in gene mapper with a positive control and all samples were confirmed to be negative.

#### **4.4 Discussion**

Our current understanding of the CDCV and MRV hypotheses suggests that they coexist on the spectrum of complex disease etiology, with different variants in the same gene manifesting increased disease risk or disease causality. This concept of graded risk is challenging to approach from an experimental point of view, as certain techniques are more suitable to the discovery of common variants (i.e. genotyping) while others are more appropriate for the identification of rare variants (i.e. WES, WGS). Previous studies assessing the heritability of PD have been fruitful in regards to our progress in identifying both rare and common PD variants. Notably, the heritable component of PD has been estimated to be around 30%.<sup>49</sup> However, as we can only account for a small percentage of this through known PD risk associated variants, this suggests there are significantly more PD risk and causal variants that have yet to be identified. In 2014, Kara et. al assessed PD risk loci in several Greek populations by genotyping known risk alleles in both cases and controls.<sup>333</sup> This data revealed that the PD risk genes in Northern European and American populations are likewise representative of several Greek populations; thus, it follows that the missing heritability underlying PD risk genes in these populations is pertinent to the Greek population as well. As only 1.27% of known risk loci were determined to account for disease among several Greek cohorts, this suggests that the genetic etiology underlying PD in Greek populations remains largely unknown.

One of the critical issues when pursuing genetic analyses in heterogeneous populations is the inability to differentiate between normal inter-population variability

with true risk variants. Even within the country of Greece, which has been described as a crossroad between Europe, Africa and the Middle East, there is substantial genetic heterogeneity due to the ebb and flow of migratory populations for many thousands of years.<sup>334</sup> Despite the vast genetic pool among individuals in Greece, a few regions have maintained a history of extreme isolation, rendering them significantly more genetically homogenous. Analogous to the long isolated Finnish populations, which has been instrumental in the identification of genes underlying ALS, the Greek Rapsani village, located in the foothills of Mount Olympus, where PD appears to cluster in a familial fashion, may hold the key to the identification of a novel genetic cause of disease. While these populations exist throughout the world, the ability to observe distinct phenotypes and obtain pedigrees from these populations is a very challenging task, requiring astute observations, record keeping skills, and tenacious drive to embark on a longitudinal study. As we have been fortunate to work with neurologist, Dr. Georgia Xiromerisiou, who has undertaken this impressive feat for the Rapsani village population, we have pursued comprehensive genetic analyses on many of these individuals.

While we were not able to obtain any parent and child sample duos, the ability to acquire several sets of siblings, first cousins, aunts and uncles was invaluable. As we had no information regarding risk and/or causal variants in any known genes, nor a clearly identified pattern of inheritance, we used several laboratory methodologies to test both CDCV and MRV hypotheses.

By generating extensive genotyping typing data in a family pedigree, the goal is to identify long regions of IBS that are shared among affected relatives. Using Plink we were able to compare observed and expected numbers of homozygosity and inbreeding



coefficients within each sample to confirm a lack of consanguinity among family members. By comparing the most distantly related affected individuals in the pedigree and identifying identical IBS regions, this likely suggests IBD due to common ancestry, which may ultimately explain disease etiology.<sup>1</sup> Hence, by genotyping all individuals in several families, we aimed to identify long IBS regions among affected members in different Rapsani families, which would suggest both a locus and plausible cause of disease. However, our Plink and GERMLINE analyses to detect runs of homozygosity and shared haplotypes were unable to identify such a region. Plausibly, given the extended period of isolation of this population, an IBD mutation-containing region could be relatively small in size, broken by recombination over generations. Thus it is feasible that our minimal region of 1Mb may have been too large. However, because our hypothesis was that the underlying cause was a single genetic mutation we pursued methodologies that would allow us to detect rare and novel variants in the Rapsani Greek village.

Using high quality WES data generated in the laboratory, we first wanted to confirm the absence of PD causing variants in all affected individuals before proceeding with our filtering analysis. Once we confirmed this, we took different approaches to analyze the data in attempt to balance the possibility of filtering out critical variants (with harsh filtering) while obtaining a candidate gene list that was feasible to analyze (with liberal filtering). Despite using both approaches, we were unable to identify a rare variant that was shared by all affected members and absent in unaffected members that could be confirmed by Sanger sequencing. While we did confirm some positive results, such as a rare variant in *AGAPI* present in 4 affected individuals, and another in *MYOIC*

carried by 5 affected individuals, the fact that other affected samples did not carry these variants did not fit with our original hypothesis that disease in this isolate was likely to be caused by a single mutation.

In order to address the possibility that a rare and novel mutation was inadequately captured using WES, we pursued WGS. Despite obtaining very high depth and coverage of all 7 affected samples sequenced, our individual analyses of CNVs, SVs and SNP-Indels were unremarkable. One caveat to this is that the CNV data has not been investigated at great lengths; however, given the extensive candidate list, investigation of each shared CNV loss or gain without knowledge of a specific chromosome or loci is impractical at this stage. Nonetheless, as we know that duplications and triplications in *SNCA* are a cause of PD, demonstrating a gene dosage effect, further investigation of the CNV results holds promise. Further, as the age of onset of PD is quite variable in different generations of the Rapsani families, a gene dosage effect is plausible. While the SV results were likewise unremarkable in coding regions, we cannot exclude the possibility of an intronic repeat such as *C9ORF72*, as a cause of disease. Similar to the graded effect of gene dosage, a pathological repeat may vary in length in affected individuals, adhering to the concept of anticipation. Likewise, this could explain the highly variable age of onset and phenotypic severity of disease. Finally, while the SNP-Indel results revealed some possible candidate genes for individual family analyses, the extreme isolation of the village strongly suggests the presence of a single founder mutation.

As a population under isolation since 300 BC, we hypothesized that given the uncharacteristically high prevalence of PD among the Rapsani village, a very rare coding

mutation is likely to explain disease in such a genetically homogenous population. From our extensive genotyping, WES and WGS analyses, we cannot rule out this possibility entirely, but it appears to be unlikely given our results that failed to identify disease segregating regions or mutations. We entertained the possibility of both heterozygous and homozygous modes of inheritance, and even accounted for a “pseudo-dominant” model of inheritance, which occurs when there are heterozygous and homozygous mating patterns. While both individuals may be affected, a homozygous individual may manifest disease significantly earlier in life or present with greater phenotypic severity. However, the coding data likewise did not reveal any variants that suggested this pattern of inheritance.

In viewing the data and analysis performed thus far it is useful to speculate why no single disease causing mutation has been identified and what steps can be taken to further investigate the underlying cause of disease in this population.

First, it is possible that the disease investigated here is either not of simple genetic origin or of any significant genetic origin. The former might suggest that disease is caused by a confluence of low risk variants in affected individuals. While this is a possibility, the strong familial nature of this disease would argue against this possibility, given the high degree of homogeneity in Rapsani, one might expect a large number of apparently sporadic PD cases if there were indeed a large number of low to moderate risk variants. It is, however, possible, that the disease noted here is driven by factors outside of genetics. While at face value the familiarity of disease may argue against this it is possible that an environmental risk factor is enriched within certain families in Rapsani;

however, with the changing lifestyle associated with modernization of this area, this possibility remains difficult to assess prospectively.

Another key consideration in our analysis is the concept of ascertainment bias; in the context of the Rapsani village, we have sampled a very small subset of the estimated ~1500 population size. While our best source of PD prevalence was obtained via personal communication with Dr. Georgia Xiromerisiou, the absence of any literature validating these approximations is not trivial. For example, while we can only hypothesize at this point, it is possible that the Rapsani population subsides primarily on a Mediterranean diet, affording individuals longer lifespans as compared to other Caucasian individuals of European descent. Ultimately, an increased prevalence of PD in the Rapsani population simply may be attributed to their longevity, as age is the greatest risk factor for the development of PD. Hence, our presumption of the strong likelihood that this PD cluster is caused by a single gene may be premature, as it is plausible that Rapsani villagers are developing sporadic forms of PD given a risk factor of advanced age. Thus, while the demographic and epidemiological information obtained about the Rapsani village is confined to the limits of our collaboration, we must acknowledge that age-adjusted prevalence is necessary to address the issue of ascertainment bias. Moreover, the ability to acquire genetic samples of other villagers and determine the cryptic relatedness and inbreeding coefficients between each other would be instrumental towards deriving accurate Rapsani population allele frequency distributions, further minimizing additional sampling bias.<sup>347</sup>

An additional possibility is that there is more than one cause of disease in this population, i.e. that some cases are a result of a single genetic mutation, and some are

sporadic in nature. Based on the segregation and sequencing data generated thus far we would posit that this is the most likely scenario. If this is true it represents a particular challenge to analysis and, absent an a priori hypothesis of the genetic and non-genetic cases, a large list of candidate variants. This is illustrated by the rather large number of variants identified when reducing the number of required mutation carriers to 6 of 7 or 5 of 7 affected family members. A plausible approach to addressing this possibility is to screen the genes containing such mutations in additional cohorts of PD cases for which there exists extensive sequence data. This is certainly a credible approach that can be pursued as such data becomes available.

As we have acquired such thorough clinical histories a plausible approach to variant identification centers on grouping affected individuals based on phenotype and looking for disease segregating mutations within each distinct group. As above this represents a significant problem by reducing power (and an inverse increase in the number of plausible variants). Also, it is worth noting that thus far, even within families with a single mutation (for e.g. *LRRK2* pG2019S), significant phenotypic variability can be observed.

While it is clear that there will be significant challenges in further analyzing the Rapsani PD cases, the added information garnered from identifying novel genetic causes of disease makes this a worthy endeavor and based on what we have seen in this village the prevalence of PD in the Rapsani village is unlikely to be explained by chance alone.

## 5 Conclusions and future directions

The study of the genetic architecture underlying monogenic and complex disorders has been advanced through application of state-of-the-art genomic technologies. The identification of both common and rare variants in several neurodegenerative diseases has provided momentum to pursue additional studies and meta-analyses. By understanding the coexistence of CDCV and MRV hypotheses, which define the cornerstones of the PRL paradigm, our integrated approach to study genetics underlying complex disease etiology has been lucrative. This is not only in regard to the discovery of graded risk factors associated with disease, but additionally for the identification of genomic landscape that does not merit further investigation. Principally, this provides the scientific community with a guide of where to direct subsequent analyses, as a map would provide geographic and topographic information. Likewise, the results of negative experiments and those unable to withstand replication are critical to hinder further investment of time, effort, finances and other resources into fruitless endeavors; hence, we learn which geographic territories on the map to no longer explore.

In my thesis, I aimed to use these approaches towards the study of genetics in two different neurological disorders. The underlying rationale of the specific genomic technologies applied and data analysis strategies were tailored to each study. This comprised several factors including: prevalence of disease, previously known genetic information about disease, relationship between affected samples (if any), among others. By integrating this information with respect to each project, I was able to accomplish my chief objectives.

### 5.1.1 Chapter 2 overview

In my second chapter, I focused on estimating the heritability of a rare and understudied disease, MSA. This was accomplished by using the previous MSA GWA study data obtained in 2011 including >900 samples in total. However, noting that pathological confirmation is required for a definitive diagnosis, some of these samples were only assessed in a clinical setting. Despite several GWA study hits with a p-value <  $1 \times 10^{-6}$ , none of these loci were deemed statistically significant at a genome-wide level. While unraveling the identity of genetic risk variants in MSA was not feasible using the current GWA results, prior investigations in PD and ALS suggested that these data can be effectively used to quantify heritability through polygenic additive inheritance analyses.<sup>49,329</sup> Furthermore, as the ease of obtaining several thousands of MSA samples, likely needed to obtain requisite statistical power for a MSA GWA study, is extremely challenging, this represented a practical approach that would provide information on whether it would be important to use such samples for GWA. By providing a heritability estimate of an apparently sporadic disease, we hoped to glean insight on the following: first, to determine if MSA is indeed heritable by quantifying this amount. Secondly, if substantial heritability was estimated, pursuing further investigation of the particular loci harboring common variants that revealed association with disease. These variants, which we hypothesized could be either protective or deleterious towards the development of MSA, would nonetheless yield insight into the pathogenesis of disease.

Despite the estimated heritability of common variants underlying other neurodegenerative disorders (i.e. PD, ALS) residing between 20-30%, our estimate was only in the 2.09-6.65% range after imputation for MSA. Further, given the estimated

misdiagnosis rate of MSA of 14%, which we presume is most commonly misdiagnosed PD, we used a Bayesian approach to calculate the expected heritability due to misdiagnosis, which we measured at a range of 1.00-4.19%. As these ranges substantially overlap, we hypothesized that all MSA heritability estimated in this study could in theory be attributed exclusively to heritability stemming from samples of a non-MSA origin.

While we acknowledged several limitations in genomic technologies used to quantify our estimate of MSA heritability, our results suggested that common variation is unlikely to play a role in genetic etiology of MSA. However, as the genotyping technology used was not amenable for rare variant detection, this was the next practical area to investigate the genetic architecture of MSA. Hence, we proceeded to MSA exome sequencing and analysis in Chapter 3.

### **5.1.2 Chapter 3 overview**

We pursued MSA WES with the goal of identifying a list of genes as evidence based candidates for association with MSA. By filtering based on MAF, we initiated the discovery phase of rare variant identification, as we believed this was the most logical step subsequent to our heritability analysis. While maximizing cohort sample size, with at least half of our samples receiving a confirmed pathological diagnosis of MSA, we also incorporated multiple analytical approaches to generate the most comprehensive list of candidate genes and variants for future independent validation and replication in the MSA scientific community.

We performed analyses for alignment and base calling on local software as well as on GoogleGenomics. This was not only advantageous by generating an extensive set of results from both pipelines, but furthermore the ability to compare and contrast



overlapping results between them was valuable regarding both interpretation and application to future WES projects. Using local alignment, we generated a list of 13 candidate genes, all with novel variants confirmed by Sanger sequencing amongst our combined clinically diagnosed and pathologically confirmed MSA case cohort. By incorporating our data into the Googlegenome pipeline, which demonstrates increased sensitivity to detect rare variants and thus generates higher quality data, we could compare these findings with our previous list of locally derived candidates.

In addition to apparent gene burden harbored by *LRRK2*, several other functionally relevant genes revealed significant burdens including *SLC44A5*, *GLIPR1* and *CASP8AP2*. Each of the genes likewise demonstrated several significant single variants, many of which were non-synonymous. Among these, those single variants present exclusively in our case population (*SLC44A5*, *GLIPR1*) failed to validate while those carried by a disproportionately higher number of cases than controls (*CASP8AP2*) were confirmed with Sanger sequencing. Resonating with the results of *LRRK2* p.G2385R, we must be cautious in our interpretation, however, by obtaining results from local and Googlegenome pipelines, we have successfully generated a list of candidate genes which should be prioritized in a subsequent combined validation replication stage of unraveling MSA genetic architecture. This will likely entail both genotyping and resequencing approaches. However, by commencing the discovery phase (our chief objective) and generating hypothesis generated results, we have laid the foundation towards the dissection of MSA genetic etiology.

### 5.1.3 Chapter 4 overview

In chapter 4, we aimed to unravel the genetic architecture of PD observed in a Greek village that has sustained a high degree of genetic isolation for the last several centuries. In order to investigate the genetic landscape of the Rapsani families, we used several technological and analytical approaches including genotyping, homozygosity mapping, WES and WGS.

In an effort to identify long IBS regions among affected members in different Rapsani families, which would suggest both a locus and plausible cause of disease, we pursued genotyping of all individuals among the 5 families. Nonetheless, we were unsuccessful at detecting any runs of homozygosity and shared haplotypes using Plink and GERMLINE analyses.

Using WES, we were able to exclude all known PD causal variants in all affected samples, moving on to a MAF based filtering pipeline. In an effort to balance the possibility of filtering out crucial variants (with harsh filtering) while obtaining a candidate gene list that was feasible to analyze (with liberal filtering), we incorporated several different strategies for data analysis. Despite our versatility in this approach, we were unable to identify and validate a rare variant that was shared by all affected members and absent in unaffected members.

As WES coverage, depth and quality is inferior to that of WGS, we pursued the latter with 7 affected samples from 2 different Rapsani families. Results of coding data in the form of SNPs and SVs were carefully scrutinized but did not yield any notable findings. In the CNV data obtained for all sequenced individuals, we could identify

several shared CNV losses and gains among all 7 affected individuals. However, without any chromosome or locus to serve as a guide, we did not investigate these shared CNVs in great detail.

Finally, homozygosity mapping results of genotyping, WES SNP data and WGS SV and SNP-Indel data were all unremarkable. Thus, while a rare coding variant shared by all members is unlikely given our extensive analyses, we cannot exclude this possibility entirely. Given our absence of interesting findings, we have considered plausible explanations including: a disease-causing CNV manifesting a gene dosage effect, or a pathological intronic repeat region varying in length, adhering to anticipation. Furthermore, we have also acknowledged the possibility of a confluence of low risk factors among affected individuals or simply that disease may not be attributed to genetic etiology (perhaps through environmental factors); however, given the incidence of PD and genetic homogeneity of Rapsani village, these explanations seem improbable. Finally, we recognize there may be multiple causes of disease within the village, contributing to both familial and sporadic cases. Thus, a next logical step involves subsetting affected individuals based on phenotype and searching for disease segregating rare variants within each distinct group.

We acknowledge that there will be significant obstacles in further analyzing the Rapsani PD cases; however, given the wealth of clinical information and incidence of PD in the village, we believe this venture is well warranted towards the identification of a novel genetic PD etiology. As we have already generated and analyzed abundant segregation and sequencing data, we have laid the groundwork to elucidate the Greek Rapsani village PD enigma.

#### **5.1.4 Final thoughts and future directions**

Our extensive genetic analyses for underlying etiology in both MSA and PD have been feasible from the collaboration between myself and others at NIH, UCL and several other institutions. In order to pursue subsequent studies in both of these domains, it is essential that we expand our international collaboration networks to maximize our ability to detect novel variants. For such a rare and understudied disease such as MSA, increasing collaboration and expanding sample sizes will be instrumental in our ability to detect, validate and replicate findings. Our heritability analyses suggest we focus on rare variants.

With respect to PD in Rapsani, subsequent analyses in the form of CNVs and intronic regions will be pursued, as well as more in-depth individual family candidate analyses. This too, will require extensive collaboration; as we may identify several more candidates, we will need to mine for such variants and genes among an extensive global PD database. Given the prevalence of disease and homogeneity of the Rapsani village, we must cast our net wide to identify a shared pathological variant between these familial cases by assessing all plausible PD variant and gene candidates currently known in the scientific community.

Most importantly, while the studies discussed in this thesis have guided us to explore certain paths, we must maintain an open-mind in our future analyses. In regards to MSA, while we believe common variation is unlikely to play a substantial role, and will focus on rare and novel variants, we must continue to acknowledge our limitations in derivation of such data and not eliminate common variation from our umbrella of hypotheses.

Finally, regarding the Greek Rapsani village, many factors are highly suggestive of a single pathogenic mutation that we were unable to identify thus far. While this is overwhelmingly likely, we recognize the role of the PRL hypothesis and must consider all graded risk variants in our subsequent analyses.

As MSA and PD are severely debilitating and fatal neurodegenerative diseases, both of these warrant significant scientific investigation. The fact that current PD therapies target genes identified by genetic analyses (i.e. *LRRK2*) is a direct testament to the correlation between genetic discovery with clinical trials and therapies. Thus, as we continue on our journey to explore the genetic architecture for both diseases, we must not consider failed validation or replication of results as setbacks, but rather as momentum and motivation to drive us forward in our understanding of disease etiology.

## 6 Acknowledgements

DNA and brain samples: We used DNA samples and phenotype data from the NINDS Human Genetics Resource Center DNA and Cell Line Repository at Coriell (Newark, NJ, USA; <http://ccr.coriell.org/ninds>), and we would like to thank the patients and the submitters who contributed samples to this repository. Human tissue was kindly obtained from the Queen Square Brain Bank (London, UK), the institute of Psychiatry Brain Bank, King's College (London, UK), the UK Parkinson's disease tissue bank at Imperial College (London, UK), Newcastle Brain Tissue Resource at Newcastle University (Newcastle, UK), and the Manchester Brain Bank at the University of Manchester (Manchester, UK), Jacksonville Brain Bank for Alzheimer's, Parkinson's and Related Disorders at the Mayo Clinic (Jacksonville, FL, USA), the NICHD Brain and Tissue Bank for Developmental Disorders at the University of Maryland (Baltimore, MD, USA), from the New York Brain Bank of the Taub Institute at Columbia University (New York, NY, USA), the Human Brain and Spinal Fluid Resource Center (Los Angeles, CA, USA), the Miami Brain Bank (Miami, FL, USA), the Center for Neurodegenerative Disease Research at the University of Pennsylvania (Philadelphia, PA, USA), the Harvard Brain Bank (Boston, MA, USA), the Emory University Alzheimer's Disease Research Center Brain Bank (Atlanta, GA, USA), Neurobiobank Muenchen at the Ludwig Maximilians-Universitat (Munich, Germany), Brain Bank Center Wurzburg (Wurzburg, Germany), at the Netherlands Brain Bank at the Netherlands Institute for Neuroscience (Amsterdam, Netherlands), and the Neurological Tissue Bank at the University of Barcelona (Barcelona, Spain).

Samples derived from the Rapsani Greek village were obtained by neurologist, Dr. Georgia Xiromerisiou.

This work was supported supported in part by the Intramural Research Program of the National Institute on Aging, National Institutes of Health, Department of Health and Human services.

This study utilized the high-performance computational capabilities of the Biowulf Linux cluster at the National Institutes of Health, Bethesda, MD. (<http://biowulf.nih.gov>)

## 7 References

1. Manolio, T. A. *et al.* Finding the missing heritability of complex diseases. *Nature* **461**, 747–753 (2009).
2. Singleton, A. B., Hardy, J., Traynor, B. J. & Houlden, H. Towards a complete resolution of the genetic architecture of disease. *Trends Genet. TIG* **26**, 438–442 (2010).
3. Lee, J.-M. *et al.*, PREDICT-HD study of the Huntington Study Group (HSG), Landwehrmeyer, G. B., REGISTRY study of the European Huntington's Disease Network, Myers, R. H., HD-MAPS Study Group, and MacDonald, M. E. & Gusella, J. F., COHORT study of the HSG. CAG repeat expansion in Huntington disease determines age at onset in a fully dominant fashion. *Neurology* **78**, 690–695 (2012).
4. Weber, D. & Helentjaris, T. Mapping RFLP loci in maize using B-A translocations. *Genetics* **121**, 583–590 (1989).
5. Venter, J. C. *et al.* The sequence of the human genome. *Science* **291**, 1304–1351 (2001).
6. Lander, E. S. & Botstein, D. Homozygosity mapping: a way to map human recessive traits with the DNA of inbred children. *Science* **236**, 1567–1570 (1987).
7. Bras, J., Guerreiro, R. & Hardy, J. Use of next-generation sequencing and other whole-genome strategies to dissect neurological disease. *Nat. Rev. Neurosci.* **13**, 453–464 (2012).
8. Keller, M. C. *et al.*, Schizophrenia Psychiatric Genome-Wide Association Study Consortium. Runs of homozygosity implicate autozygosity as a schizophrenia risk factor. *PLoS Genet.* **8**, e1002656 (2012).

9. Risch, N. & Merikangas, K. The future of genetic studies of complex human diseases. *Science* **273**, 1516–1517 (1996).
10. Botstein, D. & Risch, N. Discovering genotypes underlying human phenotypes: past successes for mendelian disease, future approaches for complex disease. *Nat. Genet.* **33 Suppl**, 228–237 (2003).
11. Choi, M. *et al.* Genetic diagnosis by whole exome capture and massively parallel DNA sequencing. *Proc. Natl. Acad. Sci. U. S. A.* **106**, 19096–19101 (2009).
12. Hodges, E. *et al.* Genome-wide in situ exon capture for selective resequencing. *Nat. Genet.* **39**, 1522–1527 (2007).
13. Antonarakis, S. E., Chakravarti, A., Cohen, J. C. & Hardy, J. Mendelian disorders and multifactorial traits: the big divide or one for all? *Nat. Rev. Genet.* **11**, 380–384 (2010).
14. Koboldt, D. C. *et al.* VarScan 2: somatic mutation and copy number alteration discovery in cancer by exome sequencing. *Genome Res.* **22**, 568–576 (2012).
15. Bhatia, K. & Bras, J. with Kun-Rodrigues, C. *et al.*, International Parkinson's Disease Genomics Consortium (IPDGC). A systematic screening to identify de novo mutations causing sporadic early-onset Parkinson's disease. *Hum. Mol. Genet.* **24**, 6711–6720 (2015).
16. Singleton, A. & Hardy, J. A generalizable hypothesis for the genetic architecture of disease: pleomorphic risk loci. *Hum. Mol. Genet.* **20**, R158–162 (2011).
17. Gahl, W. A. *et al.* The National Institutes of Health Undiagnosed Diseases Program: insights into rare diseases. *Genet. Med. Off. J. Am. Coll. Med. Genet.* **14**, 51–59 (2012).



18. Bamshad, M. J. with Chong, J. X. *et al.*, Centers for Mendelian Genomics. The Genetic Basis of Mendelian Phenotypes: Discoveries, Challenges, and Opportunities. *Am. J. Hum. Genet.* **97**, 199–215 (2015).
19. Abecasis, G. R. *et al.* with 1000 Genomes Project Consortium. A map of human genome variation from population-scale sequencing. *Nature* **467**, 1061–1073 (2010).
20. Schossig, A. *et al.* Mutations in ROGDI Cause Kohlschütter-Tönz Syndrome. *Am. J. Hum. Genet.* **90**, 701–707 (2012).
21. De Souza, C. M. *et al.* Kohlschütter-Tönz syndrome in siblings without ROGDI mutation. *Oral Health Dent. Manag.* **13**, 728–730 (2014).
22. Huckert, M. *et al.* A Novel Mutation in the ROGDI Gene in a Patient with Kohlschütter-Tönz Syndrome. *Mol. Syndromol.* **5**, 293–298 (2014).
23. Renton, A. E. *et al.* A hexanucleotide repeat expansion in C9ORF72 is the cause of chromosome 9p21-linked ALS-FTD. *Neuron* **72**, 257–268 (2011).
24. Kirby, A. *et al.* Mutations causing medullary cystic kidney disease type 1 lie in a large VNTR in MUC1 missed by massively parallel sequencing. *Nat. Genet.* **45**, 299–303 (2013).
25. Sebat, J. *et al.* Strong association of de novo copy number mutations with autism. *Science* **316**, 445–449 (2007).
26. Ledbetter, D. H. *et al.* with Sanders, S. J. *et al.*, Autism Sequencing Consortium. Insights into Autism Spectrum Disorder Genomic Architecture and Biology from 71 Risk Loci. *Neuron* **87**, 1215–1233 (2015).

27. Strittmatter, W. J. & Roses, A. D. Apolipoprotein E and Alzheimer's disease. *Annu. Rev. Neurosci.* **19**, 53–77 (1996).
28. Patnala, R., Clements, J. & Batra, J. Candidate gene association studies: a comprehensive guide to useful in silico tools. *BMC Genet.* **14**, 39 (2013).
29. International HapMap Consortium. The International HapMap Project. *Nature* **426**, 789–796 (2003).
30. Li, Y., Willer, C., Sanna, S. & Abecasis, G. Genotype imputation. *Annu. Rev. Genomics Hum. Genet.* **10**, 387–406 (2009).
31. Pritchard, J. K., Stephens, M., Rosenberg, N. A. & Donnelly, P. Association mapping in structured populations. *Am. J. Hum. Genet.* **67**, 170–181 (2000).
32. Young, J. H. *et al.* Differential susceptibility to hypertension is due to selection during the out-of-Africa expansion. *PLoS Genet.* **1**, e82 (2005).
33. Klein, R. J. *et al.* Complement factor H polymorphism in age-related macular degeneration. *Science* **308**, 385–389 (2005).
34. Johnson, A. D. & O'Donnell, C. J. An open access database of genome-wide association results. *BMC Med. Genet.* **10**, 6 (2009).
35. Visscher, P. M., Brown, M. A., McCarthy, M. I. & Yang, J. Five years of GWAS discovery. *Am. J. Hum. Genet.* **90**, 7–24 (2012).
36. Balding, D. J. A tutorial on statistical methods for population association studies. *Nat. Rev. Genet.* **7**, 781–791 (2006).
37. Simón-Sánchez, J. *et al.* Genome-wide association study reveals genetic risk underlying Parkinson's disease. *Nat. Genet.* **41**, 1308–1312 (2009).

38. Batzoglou, S. *et al.* with ENCODE Project Consortium *et al.* Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. *Nature* **447**, 799–816 (2007).
39. Wang, M.-H., Guo, M. & Shugart, Y. Y. Application of family-based association testing to assess the genotype-phenotype association involved in complex traits using single-nucleotide polymorphisms. *BMC Genet.* **6 Suppl 1**, S68 (2005).
40. Guerreiro, P. S. *et al.* LRRK2 interactions with  $\alpha$ -synuclein in Parkinson's disease brains and in cell models. *J. Mol. Med. Berl. Ger.* **91**, 513–522 (2013).
41. Dickson, S. P., Wang, K., Krantz, I., Hakonarson, H. & Goldstein, D. B. Rare variants create synthetic genome-wide associations. *PLoS Biol.* **8**, e1000294 (2010).
42. Goldstein, D. B. The importance of synthetic associations will only be resolved empirically. *PLoS Biol.* **9**, e1001008 (2011).
43. Rice, T. K., Schork, N. J. & Rao, D. C. Methods for handling multiple testing. *Adv. Genet.* **60**, 293–308 (2008).
44. Guerreiro, R. *et al.* TREM2 variants in Alzheimer's disease. *N. Engl. J. Med.* **368**, 117–127 (2013).
45. Maher, B. Personal genomes: The case of the missing heritability. *Nature* **456**, 18–21 (2008).
46. McClellan, J. & King, M.-C. Genomic analysis of mental illness: a changing landscape. *JAMA* **303**, 2523–2524 (2010).
47. Visscher, P. M., Hill, W. G. & Wray, N. R. Heritability in the genomics era-- concepts and misconceptions. *Nat. Rev. Genet.* **9**, 255–266 (2008).

48. Yang, J., Lee, S. H., Goddard, M. E. & Visscher, P. M. GCTA: a tool for genome-wide complex trait analysis. *Am. J. Hum. Genet.* **88**, 76–82 (2011).
49. Keller, M. F. *et al.*, International Parkinson's Disease Genomics Consortium (IPDGC) and Wellcome Trust Case Control Consortium 2 (WTCCC2). Using genome-wide complex trait analysis to quantify 'missing heritability' in Parkinson's disease. *Hum. Mol. Genet.* **21**, 4996–5009 (2012).
50. Lieber, D. S. *et al.* Atypical case of Wolfram syndrome revealed through targeted exome sequencing in a patient with suspected mitochondrial disease. *BMC Med. Genet.* **13**, 3 (2012).
51. Ionita-Laza, I. *et al.* Finding disease variants in Mendelian disorders by using sequence data: methods and applications. *Am. J. Hum. Genet.* **89**, 701–712 (2011).
52. Gonzaga-Jauregui, C., Lupski, J. R. & Gibbs, R. A. Human genome sequencing in health and disease. *Annu. Rev. Med.* **63**, 35–61 (2012).
53. Fogel, B. L., Clark, M. C. & Geschwind, D. H. The neurogenetics of atypical parkinsonian disorders. *Semin. Neurol.* **34**, 217–224 (2014).
54. Schork, N. J., Murray, S. S., Frazer, K. A. & Topol, E. J. Common vs. rare allele hypotheses for complex diseases. *Curr. Opin. Genet. Dev.* **19**, 212–219 (2009).
55. Koboldt, D. C., Steinberg, K. M., Larson, D. E., Wilson, R. K. & Mardis, E. R. The next-generation sequencing revolution and its impact on genomics. *Cell* **155**, 27–38 (2013).
56. Galassi, G. *et al.* with Johnson, J. O. *et al.*, ITALSGEN Consortium. Exome sequencing reveals VCP mutations as a cause of familial ALS. *Neuron* **68**, 857–864 (2010).

57. Drepper, C. *et al.* with Johnson, J. O. *et al.*, ITALSGEN Consortium. Mutations in the Matrin 3 gene cause familial amyotrophic lateral sclerosis. *Nat. Neurosci.* **17**, 664–666 (2014).
58. Zimprich, A. *et al.* A mutation in VPS35, encoding a subunit of the retromer complex, causes late-onset Parkinson disease. *Am. J. Hum. Genet.* **89**, 168–175 (2011).
59. Vilariño-Güell, C. *et al.* VPS35 mutations in Parkinson disease. *Am. J. Hum. Genet.* **89**, 162–167 (2011).
60. Guerreiro, R. & Hardy, J. TREM2 and neurodegenerative disease. *N. Engl. J. Med.* **369**, 1569–1570 (2013).
61. Lupski, J. R. *et al.* Exome sequencing resolves apparent incidental findings and reveals further complexity of SH3TC2 variant alleles causing Charcot-Marie-Tooth neuropathy. *Genome Med.* **5**, 57 (2013).
62. Spatola, M. & Wider, C. Genetics of Parkinson's disease: the yield. *Parkinsonism Relat. Disord.* **20 Suppl 1**, S35–38 (2014).
63. De Lau, L. M. L. & Breteler, M. M. B. Epidemiology of Parkinson's disease. *Lancet Neurol.* **5**, 525–535 (2006).
64. Trinh, J. & Farrer, M. Advances in the genetics of Parkinson disease. *Nat. Rev. Neurol.* **9**, 445–454 (2013).
65. Verstraeten, A., Theuns, J. & Van Broeckhoven, C. Progress in unraveling the genetic etiology of Parkinson disease in a genomic era. *Trends Genet. TIG* **31**, 140–149 (2015).

66. Chaudhuri, K. R., Healy, D. G. & Schapira, A. H. V., National Institute for Clinical Excellence. Non-motor symptoms of Parkinson's disease: diagnosis and management. *Lancet Neurol.* **5**, 235–245 (2006).
67. Dickson, D. W. Parkinson's disease and parkinsonism: neuropathology. *Cold Spring Harb. Perspect. Med.* **2**, (2012).
68. Ikeda, K., Ikeda, S., Yoshimura, T., Kato, H. & Namba, M. Idiopathic Parkinsonism with Lewy-type inclusions in cerebral cortex. A case report. *Acta Neuropathol. (Berl.)* **41**, 165–168 (1978).
69. Houlden, H. & Singleton, A. B. The genetics and neuropathology of Parkinson's disease. *Acta Neuropathol. (Berl.)* **124**, 325–338 (2012).
70. Braak, H. & Del Tredici, K. Neuroanatomy and pathology of sporadic Parkinson's disease. *Adv. Anat. Embryol. Cell Biol.* **201**, 1–119 (2009).
71. Fujishiro, H. *et al.* Validation of the neuropathologic criteria of the third consortium for dementia with Lewy bodies for prospectively diagnosed cases. *J. Neuropathol. Exp. Neurol.* **67**, 649–656 (2008).
72. Klingelhoefer, L. & Reichmann, H. Pathogenesis of Parkinson disease--the gut-brain axis and environmental factors. *Nat. Rev. Neurol.* **11**, 625–636 (2015).
73. Healy, D. G. *et al.*, International LRRK2 Consortium. Phenotype, genotype, and worldwide genetic penetrance of LRRK2-associated Parkinson's disease: a case-control study. *Lancet Neurol.* **7**, 583–590 (2008).
74. Ross, O. A., Vilariño-Güell, C., Wszolek, Z. K., Farrer, M. J. & Dickson, D. W. Reply to: SNCA variants are associated with increased risk of multiple system atrophy. *Ann. Neurol.* **67**, 414–415 (2010).

75. Ross, O. A. *et al.*, Genetic Epidemiology Of Parkinson's Disease (GEO-PD) Consortium. Association of LRRK2 exonic variants with susceptibility to Parkinson's disease: a case-control study. *Lancet Neurol.* **10**, 898–908 (2011).
76. Bardien, S., Lesage, S., Brice, A. & Carr, J. Genetic characteristics of leucine-rich repeat kinase 2 (LRRK2) associated Parkinson's disease. *Parkinsonism Relat. Disord.* **17**, 501–508 (2011).
77. Lesage, S. *et al.*, French Parkinson's Disease Genetics Study Group. LRRK2 haplotype analyses in European and North African families with Parkinson disease: a common founder for the G2019S mutation dating from the 13th century. *Am. J. Hum. Genet.* **77**, 330–332 (2005).
78. Hasegawa, K. *et al.* Familial parkinsonism: study of original Sagamihara PARK8 (I2020T) kindred with variable clinicopathologic outcomes. *Parkinsonism Relat. Disord.* **15**, 300–306 (2009).
79. Zimprich, A. *et al.* Mutations in LRRK2 cause autosomal-dominant parkinsonism with pleomorphic pathology. *Neuron* **44**, 601–607 (2004).
80. Khan, N. L. *et al.* Mutations in the gene LRRK2 encoding dardarin (PARK8) cause familial Parkinson's disease: clinical, pathological, olfactory and functional imaging and genetic data. *Brain J. Neurol.* **128**, 2786–2796 (2005).
81. Gaig, C. *et al.* G2019S LRRK2 mutation causing Parkinson's disease without Lewy bodies. *J. Neurol. Neurosurg. Psychiatry* **78**, 626–628 (2007).
82. Ross, O. A. *et al.* Lrrk2 and Lewy body disease. *Ann. Neurol.* **59**, 388–393 (2006).
83. Ujiiie, S. *et al.* LRRK2 I2020T mutation is associated with tau pathology. *Parkinsonism Relat. Disord.* **18**, 819–823 (2012).

84. Lewis, P. A. & Alessi, D. R. Deciphering the function of leucine-rich repeat kinase 2 and targeting its dysfunction in disease. *Biochem. Soc. Trans.* **40**, 1039–1041 (2012).
85. Plun-Favreau, H., Lewis, P. A., Hardy, J., Martins, L. M. & Wood, N. W. Cancer and neurodegeneration: between the devil and the deep blue sea. *PLoS Genet.* **6**, e1001257 (2010).
86. Cardoso, C. C., Pereira, A. C., de Sales Marques, C. & Moraes, M. O. Leprosy susceptibility: genetic variations regulate innate and adaptive immunity, and disease outcome. *Future Microbiol.* **6**, 533–549 (2011).
87. Saunders-Pullman, R. *et al.* LRRK2 G2019S mutations are associated with an increased cancer risk in Parkinson disease. *Mov. Disord. Off. J. Mov. Disord. Soc.* **25**, 2536–2541 (2010).
88. Zhang, F.-R. *et al.* Genomewide association study of leprosy. *N. Engl. J. Med.* **361**, 2609–2618 (2009).
89. Hernandez, D. *et al.* The dardarin G 2019 S mutation is a common cause of Parkinson's disease but not other neurodegenerative diseases. *Neurosci. Lett.* **389**, 137–139 (2005).
90. Winner, B. *et al.* Adult neurogenesis and neurite outgrowth are impaired in LRRK2 G2019S mice. *Neurobiol. Dis.* **41**, 706–716 (2011).
91. Tong, Y. *et al.* Loss of leucine-rich repeat kinase 2 causes age-dependent bi-phasic alterations of the autophagy pathway. *Mol. Neurodegener.* **7**, 2 (2012).
92. Piccoli, G. *et al.* LRRK2 controls synaptic vesicle storage and mobilization within the recycling pool. *J. Neurosci. Off. J. Soc. Neurosci.* **31**, 2225–2237 (2011).



93. Polymeropoulos, M. H. *et al.* Mutation in the alpha-synuclein gene identified in families with Parkinson's disease. *Science* **276**, 2045–2047 (1997).
94. Singleton, A. B. *et al.* alpha-Synuclein locus triplication causes Parkinson's disease. *Science* **302**, 841 (2003).
95. Young, P. *et al.* with Lill, C. M. *et al.*, 23andMe Genetic Epidemiology of Parkinson's Disease Consortium, International Parkinson's Disease Genomics Consortium, Parkinson's Disease GWAS Consortium, and Wellcome Trust Case Control Consortium 2). Comprehensive research synopsis and systematic meta-analyses in Parkinson's disease genetics: The PDGene database. *PLoS Genet.* **8**, e1002548 (2012).
96. Zarranz, J. J. *et al.* The new mutation, E46K, of alpha-synuclein causes Parkinson and Lewy body dementia. *Ann. Neurol.* **55**, 164–173 (2004).
97. Farrer, M. *et al.* Comparison of kindreds with parkinsonism and alpha-synuclein genomic multiplications. *Ann. Neurol.* **55**, 174–179 (2004).
98. Golbe, L. I. The genetics of Parkinson's disease: a reconsideration. *Neurology* **40**, suppl 7–14; discussion 14–16 (1990).
99. Singleton, A. & Gwinn-Hardy, K. Parkinson's disease and dementia with Lewy bodies: a difference in dose? *Lancet Lond. Engl.* **364**, 1105–1107 (2004).
100. Conway, K. A., Harper, J. D. & Lansbury, P. T. Fibrils formed in vitro from alpha-synuclein and two mutant forms linked to Parkinson's disease are typical amyloid. *Biochemistry (Mosc.)* **39**, 2552–2563 (2000).

101. Baba, M. *et al.* Aggregation of alpha-synuclein in Lewy bodies of sporadic Parkinson's disease and dementia with Lewy bodies. *Am. J. Pathol.* **152**, 879–884 (1998).
102. Desplats, P. *et al.* Inclusion formation and neuronal cell death through neuron-to-neuron transmission of alpha-synuclein. *Proc. Natl. Acad. Sci. U. S. A.* **106**, 13010–13015 (2009).
103. Westphal, C. H. & Chandra, S. S. Monomeric synucleins generate membrane curvature. *J. Biol. Chem.* **288**, 1829–1840 (2013).
104. Stefanis, L.  $\alpha$ -Synuclein in Parkinson's disease. *Cold Spring Harb. Perspect. Med.* **2**, a009399 (2012).
105. Sharma, M. *et al.*, GEOPD consortium. A multi-centre clinico-genetic analysis of the VPS35 gene in Parkinson disease indicates reduced penetrance for disease-associated variants. *J. Med. Genet.* **49**, 721–726 (2012).
106. Sheerin, U.-M. *et al.* Screening for VPS35 mutations in Parkinson's disease. *Neurobiol. Aging* **33**, 838.e1–5 (2012).
107. Lubbe, S. & Morris, H. R. Recent advances in Parkinson's disease genetics. *J. Neurol.* **261**, 259–266 (2014).
108. Seaman, M. N. J. The retromer complex - endosomal protein recycling and beyond. *J. Cell Sci.* **125**, 4693–4702 (2012).
109. Zavodszky, E., Seaman, M. N. J. & Rubinsztein, D. C. VPS35 Parkinson mutation impairs autophagy via WASH. *Cell Cycle Georget. Tex* **13**, 2155–2156 (2014).
110. Kawaguchi, Y. *et al.* CAG expansions in a novel gene for Machado-Joseph disease at chromosome 14q32.1. *Nat. Genet.* **8**, 221–228 (1994).

111. Hutton, M. *et al.* Association of missense and 5'-splice-site mutations in tau with the inherited dementia FTDP-17. *Nature* **393**, 702–705 (1998).
112. Puls, I. *et al.* Mutant dynactin in motor neuron disease. *Nat. Genet.* **33**, 455–456 (2003).
113. Farrer, M. J. *et al.* DCTN1 mutations in Perry syndrome. *Nat. Genet.* **41**, 163–165 (2009).
114. Perry, T. L. *et al.* Hereditary mental depression and Parkinsonism with taurine deficiency. *Arch. Neurol.* **32**, 108–113 (1975).
115. Hjerlind, L. E. *et al.* Dopa-responsive dystonia and early-onset Parkinson's disease in a patient with GTP cyclohydrolase I deficiency? *Mov. Disord. Off. J. Mov. Disord. Soc.* **21**, 679–682 (2006).
116. Lohmann, E. *et al.*, French Parkinson's Disease Genetics Study Group and European Consortium on Genetic Susceptibility in Parkinson's Disease. How much phenotypic variation can be attributed to parkin genotype? *Ann. Neurol.* **54**, 176–185 (2003).
117. Mori, H. *et al.* Pathologic and biochemical studies of juvenile parkinsonism linked to chromosome 6q. *Neurology* **51**, 890–892 (1998).
118. Farrer, M. *et al.* Lewy bodies and parkinsonism in families with parkin mutations. *Ann. Neurol.* **50**, 293–300 (2001).
119. Pramstaller, P. P. *et al.* Lewy body Parkinson's disease in a large pedigree with 77 Parkin mutation carriers. *Ann. Neurol.* **58**, 411–422 (2005).

120. Kilariski, L. L. *et al.* Systematic review and UK-based study of PARK2 (parkin), PINK1, PARK7 (DJ-1) and LRRK2 in early-onset Parkinson's disease. *Mov. Disord. Off. J. Mov. Disord. Soc.* **27**, 1522–1529 (2012).
121. Valente, E. M. *et al.* Hereditary early-onset Parkinson's disease caused by mutations in PINK1. *Science* **304**, 1158–1160 (2004).
122. Bonifati, V. *et al.*, Italian Parkinson Genetics Network. Early-onset parkinsonism associated with PINK1 mutations: frequency, genotypes, and phenotypes. *Neurology* **65**, 87–95 (2005).
123. Criscuolo, C. *et al.* PINK1 homozygous W437X mutation in a patient with apparent dominant transmission of parkinsonism. *Mov. Disord. Off. J. Mov. Disord. Soc.* **21**, 1265–1267 (2006).
124. Weng, Y.-H. *et al.* PINK1 mutation in Taiwanese early-onset parkinsonism : clinical, genetic, and dopamine transporter studies. *J. Neurol.* **254**, 1347–1355 (2007).
125. Samaranch, L. *et al.* PINK1-linked parkinsonism is associated with Lewy body pathology. *Brain J. Neurol.* **133**, 1128–1142 (2010).
126. Bras, J. *et al.* Genetic analysis implicates APOE, SNCA and suggests lysosomal dysfunction in the etiology of dementia with Lewy bodies. *Hum. Mol. Genet.* **23**, 6139–6146 (2014).
127. Ramirez, A. *et al.* Hereditary parkinsonism with dementia is caused by mutations in ATP13A2, encoding a lysosomal type 5 P-type ATPase. *Nat. Genet.* **38**, 1184–1191 (2006).

128. Paisán-Ruiz, C. *et al.* Early-onset L-dopa-responsive parkinsonism with pyramidal signs due to ATP13A2, PLA2G6, FBXO7 and spatacsin mutations. *Mov. Disord. Off. J. Mov. Disord. Soc.* **25**, 1791–1800 (2010).
129. Hampshire, D. J. *et al.* Kufor-Rakeb syndrome, pallido-pyramidal degeneration with supranuclear upgaze paresis and dementia, maps to 1p36. *J. Med. Genet.* **38**, 680–682 (2001).
130. Bonifati, V. with Di Fonzo, A. *et al.*, Italian Parkinson Genetics Network. ATP13A2 missense mutations in juvenile parkinsonism and young onset Parkinson disease. *Neurology* **68**, 1557–1562 (2007).
131. Mullin, S. & Schapira, A. The genetics of Parkinson's disease. *Br. Med. Bull.* **114**, 39–52 (2015).
132. Usenovic, M. & Krainc, D. Lysosomal dysfunction in neurodegeneration: the role of ATP13A2/PARK9. *Autophagy* **8**, 987–988 (2012).
133. Grünewald, A. *et al.* ATP13A2 mutations impair mitochondrial function in fibroblasts from patients with Kufor-Rakeb syndrome. *Neurobiol. Aging* **33**, 1843.e1–7 (2012).
134. Shojaei, S. *et al.* Genome-wide linkage analysis of a Parkinsonian-pyramidal syndrome pedigree by 500 K SNP arrays. *Am. J. Hum. Genet.* **82**, 1375–1384 (2008).
135. Paisan-Ruiz, C. *et al.* Characterization of PLA2G6 as a locus for dystonia-parkinsonism. *Ann. Neurol.* **65**, 19–23 (2009).
136. Kurian, M. A. *et al.* Phenotypic spectrum of neurodegeneration associated with mutations in the PLA2G6 gene (PLAN). *Neurology* **70**, 1623–1629 (2008).

137. Khateeb, S. *et al.* PLA2G6 mutation underlies infantile neuroaxonal dystrophy. *Am. J. Hum. Genet.* **79**, 942–948 (2006).
138. Morgan, N. V. *et al.* PLA2G6, encoding a phospholipase A2, is mutated in neurodegenerative disorders with high brain iron. *Nat. Genet.* **38**, 752–754 (2006).
139. Hayflick, S. J. Pantothenate kinase-associated neurodegeneration (formerly Hallervorden-Spatz syndrome). *J. Neurol. Sci.* **207**, 106–107 (2003).
140. Chang, C.-L. & Lin, C.-M. Eye-of-the-Tiger sign is not Pathognomonic of Pantothenate Kinase-Associated Neurodegeneration in Adult Cases. *Brain Behav.* **1**, 55–56 (2011).
141. Kruer, M. C. *et al.* Novel histopathologic findings in molecularly-confirmed pantothenate kinase-associated neurodegeneration. *Brain J. Neurol.* **134**, 947–958 (2011).
142. Edvardson, S. *et al.* A deleterious mutation in DNAJC6 encoding the neuronal-specific clathrin-uncoating co-chaperone auxilin, is associated with juvenile parkinsonism. *PloS One* **7**, e36458 (2012).
143. Krebs, C. E. *et al.* The Sac1 domain of SYNJ1 identified mutated in a family with early-onset progressive Parkinsonism with generalized seizures. *Hum. Mutat.* **34**, 1200–1207 (2013).
144. Oostra, B. A., Barone, P., Wang, J. & Bonifati, V. with Quadri, M. *et al.*, International Parkinsonism Genetics Network. Mutation in the SYNJ1 gene associated with autosomal recessive, early-onset Parkinsonism. *Hum. Mutat.* **34**, 1208–1215 (2013).

145. Korvatska, O. *et al.* Altered splicing of ATP6AP2 causes X-linked parkinsonism with spasticity (XPDS). *Hum. Mol. Genet.* **22**, 3259–3268 (2013).
146. Paisan-Ruiz, C., Nath, P., Wood, N. W., Singleton, A. & Houlden, H. Clinical heterogeneity and genotype-phenotype correlations in hereditary spastic paraplegia because of Spatacsin mutations (SPG11). *Eur. J. Neurol.* **15**, 1065–1070 (2008).
147. Paisan-Ruiz, C., Dogu, O., Yilmaz, A., Houlden, H. & Singleton, A. SPG11 mutations are common in familial cases of complicated hereditary spastic paraplegia. *Neurology* **70**, 1384–1389 (2008).
148. Dick, K. J. *et al.* Mutation of FA2H underlies a complicated form of hereditary spastic paraplegia (SPG35). *Hum. Mutat.* **31**, E1251–1260 (2010).
149. Kruer, M. C. *et al.* Defective FA2H leads to a novel form of neurodegeneration with brain iron accumulation (NBIA). *Ann. Neurol.* **68**, 611–618 (2010).
150. Deas, E., Wood, N. W. & Plun-Favreau, H. Mitophagy and Parkinson's disease: the PINK1-parkin link. *Biochim. Biophys. Acta* **1813**, 623–633 (2011).
151. Winklhofer, K. F. Parkin and mitochondrial quality control: toward assembling the puzzle. *Trends Cell Biol.* **24**, 332–341 (2014).
152. Kondapalli, C. *et al.* PINK1 is activated by mitochondrial membrane potential depolarization and stimulates Parkin E3 ligase activity by phosphorylating Serine 65. *Open Biol.* **2**, 120080 (2012).
153. Narendra, D., Tanaka, A., Suen, D.-F. & Youle, R. J. Parkin is recruited selectively to impaired mitochondria and promotes their autophagy. *J. Cell Biol.* **183**, 795–803 (2008).

154. Youle, R. J. & Narendra, D. P. Mechanisms of mitophagy. *Nat. Rev. Mol. Cell Biol.* **12**, 9–14 (2011).
155. Gegg, M. E. *et al.* Mitofusin 1 and mitofusin 2 are ubiquitinated in a PINK1/parkin-dependent manner upon induction of mitophagy. *Hum. Mol. Genet.* **19**, 4861–4870 (2010).
156. Liu, S. *et al.* Parkinson's disease-associated kinase PINK1 regulates Miro protein level and axonal transport of mitochondria. *PLoS Genet.* **8**, e1002537 (2012).
157. Wang, X. *et al.* PINK1 and Parkin target Miro for phosphorylation and degradation to arrest mitochondrial motility. *Cell* **147**, 893–906 (2011).
158. Geisler, S. *et al.* The PINK1/Parkin-mediated mitophagy is compromised by PD-associated mutations. *Autophagy* **6**, 871–878 (2010).
159. Trempe, J.-F. & Fon, E. A. Structure and Function of Parkin, PINK1, and DJ-1, the Three Musketeers of Neuroprotection. *Front. Neurol.* **4**, 38 (2013).
160. Shin, J.-H. *et al.* PARIS (ZNF746) repression of PGC-1 $\alpha$  contributes to neurodegeneration in Parkinson's disease. *Cell* **144**, 689–702 (2011).
161. Gandhi, S. *et al.* PINK1-associated Parkinson's disease is caused by neuronal vulnerability to calcium-induced cell death. *Mol. Cell* **33**, 627–638 (2009).
162. Abou-Sleiman, P. M., Healy, D. G., Quinn, N., Lees, A. J. & Wood, N. W. The role of pathogenic DJ-1 mutations in Parkinson's disease. *Ann. Neurol.* **54**, 283–286 (2003).
163. Manning-Boğ, A. B., Schüle, B. & Langston, J. W. Alpha-synuclein-glucocerebrosidase interactions in pharmacological Gaucher models: a biological



- link between Gaucher disease and parkinsonism. *Neurotoxicology* **30**, 1127–1132 (2009).
164. Cullen, V. *et al.* Acid  $\beta$ -glucosidase mutants linked to Gaucher disease, Parkinson disease, and Lewy body dementia alter  $\alpha$ -synuclein processing. *Ann. Neurol.* **69**, 940–953 (2011).
  165. Mazzulli, J. R. *et al.* Gaucher disease glucocerebrosidase and  $\alpha$ -synuclein form a bidirectional pathogenic loop in synucleinopathies. *Cell* **146**, 37–52 (2011).
  166. Schapira, A. H. *et al.* Mitochondrial complex I deficiency in Parkinson's disease. *Lancet Lond. Engl.* **1**, 1269 (1989).
  167. Orsucci, D., Caldarazzo Ienco, E., Mancuso, M. & Siciliano, G. POLG1-related and other 'mitochondrial Parkinsonisms': an overview. *J. Mol. Neurosci. MN* **44**, 17–24 (2011).
  168. Hudson, G. *et al.* Two-stage association study and meta-analysis of mitochondrial DNA variants in Parkinson disease. *Neurology* **80**, 2042–2048 (2013).
  169. Ikram, M. A. *et al.* with Nalls, M. A. *et al.*, International Parkinson's Disease Genomics Consortium (IPDGC) *et al.* Large-scale meta-analysis of genome-wide association data identifies six new risk loci for Parkinson's disease. *Nat. Genet.* **46**, 989–993 (2014).
  170. Maraganore, D. M. *et al.*, Genetic Epidemiology of Parkinson's Disease (GEO-PD) Consortium. Collaborative analysis of alpha-synuclein gene promoter variability and Parkinson disease. *JAMA* **296**, 661–670 (2006).

171. Chiba-Falek, O., Touchman, J. W. & Nussbaum, R. L. Functional analysis of intra-allelic variation at NACP-Rep1 in the alpha-synuclein gene. *Hum. Genet.* **113**, 426–431 (2003).
172. Cronin, K. D. *et al.* Expansion of the Parkinson disease-associated SNCA-Rep1 allele upregulates human alpha-synuclein in transgenic mouse brain. *Hum. Mol. Genet.* **18**, 3274–3285 (2009).
173. Bonifati, V. LRRK2 low-penetrance mutations (Gly2019Ser) and risk alleles (Gly2385Arg)-linking familial and sporadic Parkinson's disease. *Neurochem. Res.* **32**, 1700–1708 (2007).
174. Xie, C.-L. *et al.* The association between the LRRK2 G2385R variant and the risk of Parkinson's disease: a meta-analysis based on 23 case-control studies. *Neurol. Sci. Off. J. Ital. Neurol. Soc. Ital. Soc. Clin. Neurophysiol.* **35**, 1495–1504 (2014).
175. Neumann, J. *et al.* Glucocerebrosidase mutations in clinical and pathologically proven Parkinson's disease. *Brain J. Neurol.* **132**, 1783–1794 (2009).
176. Sidransky, E. *et al.* Multicenter analysis of glucocerebrosidase mutations in Parkinson's disease. *N. Engl. J. Med.* **361**, 1651–1661 (2009).
177. Tsuang, D. *et al.* GBA mutations increase risk for Lewy body disease with and without Alzheimer disease pathology. *Neurology* **79**, 1944–1950 (2012).
178. Foltynie, T. *et al.* A genome wide linkage disequilibrium screen in Parkinson's disease. *J. Neurol.* **252**, 597–602 (2005).
179. Fung, H.-C. *et al.* Genome-wide genotyping in Parkinson's disease and neurologically normal controls: first stage analysis and public release of data. *Lancet Neurol.* **5**, 911–916 (2006).

180. Evangelou, E., Maraganore, D. M. & Ioannidis, J. P. A. Meta-analysis in genome-wide association datasets: strategies and application in Parkinson disease. *PloS One* **2**, e196 (2007).
181. Farrer, M. J. *et al.* Lrrk2 G2385R is an ancestral risk factor for Parkinson's disease in Asia. *Parkinsonism Relat. Disord.* **13**, 89–92 (2007).
182. Sesar, A. *et al.* Synaptotagmin XI in Parkinson's disease: New evidence from an association study in Spain and Mexico. *J. Neurol. Sci.* **362**, 321–325 (2016).
183. Skipper, L. *et al.* Linkage disequilibrium and association of MAPT H1 in Parkinson disease. *Am. J. Hum. Genet.* **75**, 669–677 (2004).
184. Zody, M. C. *et al.* Evolutionary toggling of the MAPT 17q21.31 inversion region. *Nat. Genet.* **40**, 1076–1083 (2008).
185. Pittman, A. M. *et al.* Linkage disequilibrium fine mapping and haplotype association analysis of the tau gene in progressive supranuclear palsy and corticobasal degeneration. *J. Med. Genet.* **42**, 837–846 (2005).
186. Sundar, P. D. *et al.* Two sites in the MAPT region confer genetic risk for Guam ALS/PDC and dementia. *Hum. Mol. Genet.* **16**, 295–306 (2007).
187. Cantwell, L. B. *et al.* with Höglinger, G. U. *et al.*, PSP Genetics Study Group. Identification of common variants influencing risk of the tauopathy progressive supranuclear palsy. *Nat. Genet.* **43**, 699–705 (2011).
188. Spillantini, M. G. & Goedert, M. Tau pathology and neurodegeneration. *Lancet Neurol.* **12**, 609–622 (2013).
189. Gan-Or, Z. *et al.* The p.L302P mutation in the lysosomal enzyme gene SMPD1 is a risk factor for Parkinson disease. *Neurology* **80**, 1606–1610 (2013).

190. Tansey, M. G. & Goldberg, M. S. Neuroinflammation in Parkinson's disease: its role in neuronal death and implications for therapeutic intervention. *Neurobiol. Dis.* **37**, 510–518 (2010).
191. Hamza, T. H. *et al.* Common genetic variation in the HLA region is associated with late-onset sporadic Parkinson's disease. *Nat. Genet.* **42**, 781–785 (2010).
192. MacLeod, D. A. *et al.* RAB7L1 interacts with LRRK2 to modify intraneuronal protein sorting and Parkinson's disease risk. *Neuron* **77**, 425–439 (2013).
193. Lee, K. S. *et al.* Phosphoinositide 3-kinase-delta inhibitor reduces vascular permeability in a murine model of asthma. *J. Allergy Clin. Immunol.* **118**, 403–409 (2006).
194. Pankratz, N. *et al.*, PSG-PROGENI and GenePD Investigators, Coordinators and Molecular Genetic Laboratories. Genomewide association study for susceptibility genes contributing to familial Parkinson disease. *Hum. Genet.* **124**, 593–605 (2009).
195. Mencacci, N. E. *et al.*, International Parkinson's Disease Genomics Consortium and UCL-exomes consortium. Parkinson's disease in GTP cyclohydrolase 1 mutation carriers. *Brain J. Neurol.* **137**, 2480–2492 (2014).
196. Tanner, C. M. *et al.* Parkinson disease in twins: an etiologic study. *JAMA* **281**, 341–346 (1999).
197. Do, C. B. *et al.* Web-based genome-wide association study identifies two novel loci and a substantial genetic component for Parkinson's disease. *PLoS Genet.* **7**, e1002141 (2011).

198. Petrucci, S., Consoli, F. & Valente, E. M. Parkinson Disease Genetics: A ‘Continuum’ From Mendelian to Multifactorial Inheritance. *Curr. Mol. Med.* (2014).
199. Beck, T., Hastings, R. K., Gollapudi, S., Free, R. C. & Brookes, A. J. GWAS Central: a comprehensive resource for the comparison and interrogation of genome-wide association studies. *Eur. J. Hum. Genet. EJHG* **22**, 949–952 (2014).
200. Gasser, T. *et al.* with Simón-Sánchez, J. *et al.*, International Parkinson’s Disease Genomics Consortium and Wellcome Trust Case Control Consortium. Cooperative genome-wide analysis shows increased homozygosity in early onset Parkinson’s disease. *PloS One* **7**, e28787 (2012).
201. Ahmed, Z. *et al.* The neuropathology, pathophysiology and genetics of multiple system atrophy. *Neuropathol. Appl. Neurobiol.* **38**, 4–24 (2012).
202. Stefanova, N., Bücke, P., Duerr, S. & Wenning, G. K. Multiple system atrophy: an update. *Lancet Neurol.* **8**, 1172–1178 (2009).
203. Bower, J. H., Maraganore, D. M., McDonnell, S. K. & Rocca, W. A. Incidence of progressive supranuclear palsy and multiple system atrophy in Olmsted County, Minnesota, 1976 to 1990. *Neurology* **49**, 1284–1288 (1997).
204. Wüllner, U. *et al.* Probable multiple system atrophy in a German family. *J. Neurol. Neurosurg. Psychiatry* **75**, 924–925 (2004).
205. Osaki, Y. *et al.* Do published criteria improve clinical diagnostic accuracy in multiple system atrophy? *Neurology* **59**, 1486–1491 (2002).
206. Scholz, S. W. *et al.* SNCA variants are associated with increased risk for multiple system atrophy. *Ann. Neurol.* **65**, 610–614 (2009).

207. Gilman, S. *et al.* Second consensus statement on the diagnosis of multiple system atrophy. *Neurology* **71**, 670–676 (2008).
208. Kiely, A. P. *et al.*  $\alpha$ -Synucleinopathy associated with G51D SNCA mutation: a link between Parkinson's disease and multiple system atrophy? *Acta Neuropathol. (Berl.)* **125**, 753–769 (2013).
209. Yoshida, M. [Multiple system atrophy - synuclein and neuronal degeneration]. *Rinshō Shinkeigaku Clin. Neurol.* **51**, 838–842 (2011).
210. Ozawa, T. *et al.* The phenotype spectrum of Japanese multiple system atrophy. *J. Neurol. Neurosurg. Psychiatry* **81**, 1253–1255 (2010).
211. Nee, L. E. *et al.* Environmental-occupational risk factors and familial associations in multiple system atrophy: a preliminary investigation. *Clin. Auton. Res. Off. J. Clin. Auton. Res. Soc.* **1**, 9–13 (1991).
212. Vanacore, N. Epidemiological evidence on multiple system atrophy. *J. Neural Transm. Vienna Austria 1996* **112**, 1605–1612 (2005).
213. Wenning, G. K., Wagner, S., Daniel, S. & Quinn, N. P. Multiple system atrophy: sporadic or familial? *Lancet Lond. Engl.* **342**, 681 (1993).
214. Stamelou, M., Quinn, N. P. & Bhatia, K. P. 'Atypical' atypical parkinsonism: New genetic conditions presenting with features of progressive supranuclear palsy, corticobasal degeneration, or multiple system atrophy-A diagnostic guide. *Mov. Disord. Off. J. Mov. Disord. Soc.* (2013). doi:10.1002/mds.25509
215. Wenning, G. K. *et al.* The natural history of multiple system atrophy: a prospective European cohort study. *Lancet Neurol.* **12**, 264–274 (2013).

216. Gilman, S. *et al.* Spinocerebellar ataxia type 1 with multiple system degeneration and glial cytoplasmic inclusions. *Ann. Neurol.* **39**, 241–255 (1996).
217. Nirenberg, M. J., Libien, J., Vonsattel, J.-P. & Fahn, S. Multiple system atrophy in a patient with the spinocerebellar ataxia 3 gene mutation. *Mov. Disord. Off. J. Mov. Disord. Soc.* **22**, 251–254 (2007).
218. Huang, Y. *et al.* Anticipation of onset age in familial Parkinson's disease without SCA gene mutations. *Parkinsonism Relat. Disord.* **12**, 309–313 (2006).
219. Schöls, L., Bauer, P., Schmidt, T., Schulte, T. & Riess, O. Autosomal dominant cerebellar ataxias: clinical features, genetics, and pathogenesis. *Lancet Neurol.* **3**, 291–304 (2004).
220. Khan, N. L. *et al.* Parkinsonism and nigrostriatal dysfunction are associated with spinocerebellar ataxia type 6 (SCA6). *Mov. Disord. Off. J. Mov. Disord. Soc.* **20**, 1115–1119 (2005).
221. Kim, J.-Y. *et al.* Spinocerebellar ataxia type 17 mutation as a causative and susceptibility gene in parkinsonism. *Neurology* **72**, 1385–1389 (2009).
222. Abele, M. *et al.* The aetiology of sporadic adult-onset ataxia. *Brain J. Neurol.* **125**, 961–968 (2002).
223. Lin, I.-S., Wu, R.-M., Lee-Chen, G.-J., Shan, D.-E. & Gwinn-Hardy, K. The SCA17 phenotype can include features of MSA-C, PSP and cognitive impairment. *Parkinsonism Relat. Disord.* **13**, 246–249 (2007).
224. Kim, H.-J. *et al.* Should genetic testing for SCAs be included in the diagnostic workup for MSA? *Neurology* **83**, 1733–1738 (2014).

225. Stemberger, S., Scholz, S. W., Singleton, A. B. & Wenning, G. K. Genetic players in multiple system atrophy: unfolding the nature of the beast. *Neurobiol. Aging* **32**, 1924.e5–14 (2011).
226. Fernagut, P.-O. & Tison, F. Animal models of multiple system atrophy. *Neuroscience* **211**, 77–82 (2012).
227. Flabeau, O., Meissner, W. G. & Tison, F. Multiple system atrophy: current and future approaches to management. *Ther. Adv. Neurol. Disord.* **3**, 249–263 (2010).
228. Wenning, G. K. *et al.* What clinical features are most useful to distinguish definite multiple system atrophy from Parkinson's disease? *J. Neurol. Neurosurg. Psychiatry* **68**, 434–440 (2000).
229. Ozawa, T. *et al.* The spectrum of pathological involvement of the striatonigral and olivopontocerebellar systems in multiple system atrophy: clinicopathological correlations. *Brain J. Neurol.* **127**, 2657–2671 (2004).
230. Yoshida, M. Multiple system atrophy: alpha-synuclein and neuronal degeneration. *Neuropathol. Off. J. Jpn. Soc. Neuropathol.* **27**, 484–493 (2007).
231. Cykowski, M. D. *et al.* Expanding the spectrum of neuronal pathology in multiple system atrophy. *Brain J. Neurol.* **138**, 2293–2309 (2015).
232. Ling, H. *et al.* Minimal change multiple system atrophy: an aggressive variant? *Mov. Disord. Off. J. Mov. Disord. Soc.* **30**, 960–967 (2015).
233. Ozawa, T. *et al.* Difference in MSA phenotype distribution between populations: genetics or environment? *J. Park. Dis.* **2**, 7–18 (2012).
234. Vanacore, N. *et al.* Case-control study of multiple system atrophy. *Mov. Disord. Off. J. Mov. Disord. Soc.* **20**, 158–163 (2005).



235. Davidson, W. S., Jonas, A., Clayton, D. F. & George, J. M. Stabilization of alpha-synuclein secondary structure upon binding to synthetic membranes. *J. Biol. Chem.* **273**, 9443–9449 (1998).
236. Lee, P. H. *et al.* Serum cholesterol levels and the risk of multiple system atrophy: a case-control study. *Mov. Disord. Off. J. Mov. Disord. Soc.* **24**, 752–758 (2009).
237. Hara, K. *et al.* Multiplex families with multiple system atrophy. *Arch. Neurol.* **64**, 545–551 (2007).
238. Vidal, J.-S., Vidailhet, M., Derkinderen, P., Tzourio, C. & Alperovitch, A. Familial aggregation in atypical Parkinson's disease: a case control study in multiple system atrophy and progressive supranuclear palsy. *J. Neurol.* **257**, 1388–1393 (2010).
239. Multiple-System Atrophy Research Collaboration. Mutations in COQ2 in familial and sporadic multiple-system atrophy. *N. Engl. J. Med.* **369**, 233–244 (2013).
240. Haik, S. *et al.* Alpha-synuclein-immunoreactive deposits in human and animal prion diseases. *Acta Neuropathol. (Berl.)* **103**, 516–520 (2002).
241. Jendroska, K. *et al.* Absence of disease related prion protein in neurodegenerative disorders presenting with Parkinson's syndrome. *J. Neurol. Neurosurg. Psychiatry* **57**, 1249–1251 (1994).
242. Jeon, B. S., Farrer, M. J., Bortnick, S. F. & Korean Canadian Alliance on Parkinson's Disease and Related Disorders. Mutant COQ2 in multiple-system atrophy. *N. Engl. J. Med.* **371**, 80 (2014).
243. Sharma, M., Wenning, G., Krüger, R. & European Multiple-System Atrophy Study Group (EMSA-SG). Mutant COQ2 in multiple-system atrophy. *N. Engl. J. Med.* **371**, 80–81 (2014).

244. Schottlaender, L. V., Houlden, H. & Multiple-System Atrophy (MSA) Brain Bank Collaboration. Mutant COQ2 in multiple-system atrophy. *N. Engl. J. Med.* **371**, 81 (2014).
245. Bleasel, J. M., Wong, J. H., Halliday, G. M. & Kim, W. S. Lipid dysfunction and pathogenesis of multiple system atrophy. *Acta Neuropathol. Commun.* **2**, 15 (2014).
246. Soma, H. *et al.* Associations between multiple system atrophy and polymorphisms of SLC1A4, SQSTM1, and EIF4EBP1 genes. *Mov. Disord. Off. J. Mov. Disord. Soc.* **23**, 1161–1167 (2008).
247. Ishizawa, K. *et al.* Microglial activation parallels system degeneration in multiple system atrophy. *J. Neuropathol. Exp. Neurol.* **63**, 43–52 (2004).
248. Wyss-Coray, T. & Mucke, L. Inflammation in neurodegenerative disease--a double-edged sword. *Neuron* **35**, 419–432 (2002).
249. Combarros, O., Infante, J., Llorca, J. & Berciano, J. Interleukin-1A (-889) genetic polymorphism increases the risk of multiple system atrophy. *Mov. Disord. Off. J. Mov. Disord. Soc.* **18**, 1385–1386 (2003).
250. Nishimura, M. *et al.* Contribution of the interleukin-1beta gene polymorphism in multiple system atrophy. *Mov. Disord. Off. J. Mov. Disord. Soc.* **17**, 808–811 (2002).
251. Infante, J., Llorca, J., Berciano, J. & Combarros, O. Interleukin-8, intercellular adhesion molecule-1 and tumour necrosis factor-alpha gene polymorphisms and the risk for multiple system atrophy. *J. Neurol. Sci.* **228**, 11–13 (2005).
252. Furiya, Y. *et al.* Alpha-1-antichymotrypsin gene polymorphism and susceptibility to multiple system atrophy (MSA). *Brain Res. Mol. Brain Res.* **138**, 178–181 (2005).

253. Nishimura, M., Kuno, S., Kaji, R. & Kawakami, H. Influence of a tumor necrosis factor gene polymorphism in Japanese patients with multiple system atrophy. *Neurosci. Lett.* **374**, 218–221 (2005).
254. Shibao, C. *et al.* PRNP M129V homozygosity in multiple system atrophy vs. Parkinson's disease. *Clin. Auton. Res. Off. J. Clin. Auton. Res. Soc.* **18**, 13–19 (2008).
255. Lincoln, S. J. *et al.* Quantitative PCR-based screening of alpha-synuclein multiplication in multiple system atrophy. *Parkinsonism Relat. Disord.* **13**, 340–342 (2007).
256. Morris, H. R. *et al.* Multiple system atrophy/progressive supranuclear palsy: alpha-Synuclein, synphilin, tau, and APOE. *Neurology* **55**, 1918–1920 (2000).
257. Ozawa, T. *et al.* No mutation in the entire coding region of the alpha-synuclein gene in pathologically confirmed cases of multiple system atrophy. *Neurosci. Lett.* **270**, 110–112 (1999).
258. Ozawa, T. *et al.* The alpha-synuclein gene in multiple system atrophy. *J. Neurol. Neurosurg. Psychiatry* **77**, 464–467 (2006).
259. Ozawa, T. *et al.* Analysis of the expression level of alpha-synuclein mRNA using postmortem brain samples from pathologically confirmed cases of multiple system atrophy. *Acta Neuropathol. (Berl.)* **102**, 188–190 (2001).
260. Vogt, I. R. *et al.* Transcriptional changes in multiple system atrophy and Parkinson's disease putamen. *Exp. Neurol.* **199**, 465–478 (2006).

261. Langerveld, A. J., Mihalko, D., DeLong, C., Walburn, J. & Ide, C. F. Gene expression changes in postmortem tissue from the rostral pons of multiple system atrophy patients. *Mov. Disord. Off. J. Mov. Disord. Soc.* **22**, 766–777 (2007).
262. Al-Chalabi, A. *et al.* Genetic variants of the alpha-synuclein gene SNCA are associated with multiple system atrophy. *PloS One* **4**, e7114 (2009).
263. Satake, W. *et al.* Genome-wide association study identifies common variants at four loci as genetic risk factors for Parkinson's disease. *Nat. Genet.* **41**, 1303–1307 (2009).
264. Yun, J. Y. *et al.* SNCA variants and multiple system atrophy. *Ann. Neurol.* **67**, 554–555 (2010).
265. Guo, X. Y. *et al.* SNCA variants rs2736990 and rs356220 as risk factors for Parkinson's disease but not for amyotrophic lateral sclerosis and multiple system atrophy in a Chinese population. *Neurobiol. Aging* (2014).  
doi:10.1016/j.neurobiolaging.2014.07.014
266. Guo, X.-Y. *et al.* An association analysis of the rs1572931 polymorphism of the RAB7L1 gene in Parkinson's disease, amyotrophic lateral sclerosis and multiple system atrophy in China. *Eur. J. Neurol. Off. J. Eur. Fed. Neurol. Soc.* **21**, 1337–1343 (2014).
267. Vilarinho-Güell, C. *et al.* MAPT H1 haplotype is a risk factor for essential tremor and multiple system atrophy. *Neurology* **76**, 670–672 (2011).
268. Srulijes, K. *et al.* No association of GBA mutations and multiple system atrophy. *Eur. J. Neurol. Off. J. Eur. Fed. Neurol. Soc.* **20**, e61–62 (2013).

269. Segarane, B. *et al.* Glucocerebrosidase mutations in 108 neuropathologically confirmed cases of multiple system atrophy. *Neurology* **72**, 1185–1186 (2009).
270. Ozelius, L. J. *et al.* G2019S mutation in the leucine-rich repeat kinase 2 gene is not associated with multiple system atrophy. *Mov. Disord. Off. J. Mov. Disord. Soc.* **22**, 546–549 (2007).
271. Tan, E. K. *et al.* Analysis of 14 LRRK2 mutations in Parkinson's plus syndromes and late-onset Parkinson's disease. *Mov. Disord. Off. J. Mov. Disord. Soc.* **21**, 997–1001 (2006).
272. Heckman, M. G. *et al.* LRRK2 exonic variants and risk of multiple system atrophy. *Neurology* **83**, 2256–2261 (2014).
273. Hatano, T., Kubo, S., Sato, S. & Hattori, N. Pathogenesis of familial Parkinson's disease: new insights based on monogenic forms of Parkinson's disease. *J. Neurochem.* **111**, 1075–1093 (2009).
274. Brooks, J. A. *et al.* Mutational analysis of parkin and PINK1 in multiple system atrophy. *Neurobiol. Aging* **32**, 548.e5–7 (2011).
275. Buervenich, S. *et al.* Alcohol dehydrogenase alleles in Parkinson's disease. *Mov. Disord. Off. J. Mov. Disord. Soc.* **15**, 813–818 (2000).
276. Buervenich, S. *et al.* A rare truncating mutation in ADH1C (G78Stop) shows significant association with Parkinson disease in a large international sample. *Arch. Neurol.* **62**, 74–78 (2005).
277. Healy, D. G., Abou-Sleiman, P. M. & Wood, N. W. Genetic causes of Parkinson's disease: UCHL-1. *Cell Tissue Res.* **318**, 189–194 (2004).

278. Kim, H. S. & Lee, M. S. Frequencies of single nucleotide polymorphism in alcohol dehydrogenase7 gene in patients with multiple system atrophy and controls. *Mov. Disord. Off. J. Mov. Disord. Soc.* **18**, 1065–1067 (2003).
279. Healy, D. G. *et al.* UCHL-1 gene in multiple system atrophy: a haplotype tagging approach. *Mov. Disord. Off. J. Mov. Disord. Soc.* **20**, 1338–1343 (2005).
280. Goldman, J. S. *et al.* Multiple system atrophy and amyotrophic lateral sclerosis in a family with hexanucleotide repeat expansions in C9orf72. *JAMA Neurol.* **71**, 771–774 (2014).
281. Schottlaender, L. V., Holton, J. L. & Houlden, H. Multiple system atrophy and repeat expansions in c9orf72. *JAMA Neurol.* **71**, 1190–1191 (2014).
282. Scholz, S. W. *et al.* Multiple system atrophy is not caused by C9orf72 hexanucleotide repeat expansions. *Neurobiol. Aging* (2014).  
doi:10.1016/j.neurobiolaging.2014.08.033
283. Schottlaender, L. *et al.* The analysis of C9orf72 repeat expansions in a large series of clinically and pathologically diagnosed cases with atypical parkinsonism. *Neurobiol. Aging* doi:10.1016/j.neurobiolaging.2014.08.024
284. Sasaki, H. *et al.* Copy number loss of (src homology 2 domain containing)-transforming protein 2 (SHC2) gene: discordant loss in monozygotic twins and frequent loss in patients with multiple system atrophy. *Mol. Brain* **4**, 24 (2011).
285. Ferguson, M. C. *et al.* SHC2 gene copy number in multiple system atrophy (MSA). *Clin. Auton. Res. Off. J. Clin. Auton. Res. Soc.* **24**, 25–30 (2014).
286. Sharp, A. J. *et al.* Segmental duplications and copy-number variation in the human genome. *Am. J. Hum. Genet.* **77**, 78–88 (2005).

287. Henrichsen, C. N. *et al.* Segmental copy number variation shapes tissue transcriptomes. *Nat. Genet.* **41**, 424–429 (2009).
288. Bruder, C. E. G. *et al.* Phenotypically concordant and discordant monozygotic twins display different DNA copy-number-variation profiles. *Am. J. Hum. Genet.* **82**, 763–771 (2008).
289. Stefanova, N. *et al.* Microglial activation mediates neurodegeneration related to oligodendroglial alpha-synucleinopathy: implications for multiple system atrophy. *Mov. Disord. Off. J. Mov. Disord. Soc.* **22**, 2196–2203 (2007).
290. Block, M. L. & Hong, J.-S. Chronic microglial activation and progressive dopaminergic neurotoxicity. *Biochem. Soc. Trans.* **35**, 1127–1132 (2007).
291. Brand, A. *et al.* Membrane lipid modification by polyunsaturated fatty acids sensitizes oligodendroglial OLN-93 cells against oxidative stress and promotes up-regulation of heme oxygenase-1 (HSP32). *J. Neurochem.* **113**, 465–476 (2010).
292. Riedel, M., Goldbaum, O., Wille, M. & Richter-Landsberg, C. Membrane lipid modification by docosahexaenoic acid (DHA) promotes the formation of  $\alpha$ -synuclein inclusion bodies immunopositive for SUMO-1 in oligodendroglial cells after oxidative stress. *J. Mol. Neurosci. MN* **43**, 290–302 (2011).
293. Stefanova, N., Georgievska, B., Eriksson, H., Poewe, W. & Wenning, G. K. Myeloperoxidase inhibition ameliorates multiple system atrophy-like degeneration in a transgenic mouse model. *Neurotox. Res.* **21**, 393–404 (2012).
294. Kisos, H., Pukaß, K., Ben-Hur, T., Richter-Landsberg, C. & Sharon, R. Increased neuronal  $\alpha$ -synuclein pathology associates with its accumulation in

- oligodendrocytes in mice modeling  $\alpha$ -synucleinopathies. *PloS One* **7**, e46817 (2012).
295. Rockenstein, E. *et al.* Neuronal to oligodendroglial  $\alpha$ -synuclein redistribution in a double transgenic model of multiple system atrophy. *Neuroreport* **23**, 259–264 (2012).
296. Federoff, M., Schottlaender, L. V., Houlden, H. & Singleton, A. Multiple system atrophy: the application of genetics in understanding etiology. *Clin. Auton. Res. Off. J. Clin. Auton. Res. Soc.* **25**, 19–36 (2015).
297. Reyes, J. F. *et al.* Alpha-synuclein transfers from neurons to oligodendrocytes. *Glia* **62**, 387–398 (2014).
298. Peelaerts, W. *et al.*  $\alpha$ -Synuclein strains cause distinct synucleinopathies after local and systemic administration. *Nature* **522**, 340–344 (2015).
299. Prusiner, S. B. *et al.* Evidence for  $\alpha$ -synuclein prions causing multiple system atrophy in humans with parkinsonism. *Proc. Natl. Acad. Sci. U. S. A.* **112**, E5308–5317 (2015).
300. Wong, M. B. *et al.* SUMO-1 is associated with a subset of lysosomes in glial protein aggregate diseases. *Neurotox. Res.* **23**, 1–21 (2013).
301. Krismer, F. *et al.* Multiple system atrophy as emerging template for accelerated drug discovery in  $\alpha$ -synucleinopathies. *Parkinsonism Relat. Disord.* **20**, 793–799 (2014).
302. Nocker, M. *et al.* Progression of dopamine transporter decline in patients with the Parkinson variant of multiple system atrophy: a voxel-based analysis of [ $^{123}$ I] $\beta$ -CIT SPECT. *Eur. J. Nucl. Med. Mol. Imaging* **39**, 1012–1020 (2012).



303. Vingerhoets, F. J. *et al.* Longitudinal fluorodopa positron emission tomographic studies of the evolution of idiopathic parkinsonism. *Ann. Neurol.* **36**, 759–764 (1994).
304. Marek, K. *et al.* [123I]beta-CIT SPECT imaging assessment of the rate of Parkinson's disease progression. *Neurology* **57**, 2089–2094 (2001).
305. Wild, E. J. & Fox, N. C. Serial volumetric MRI in Parkinsonian disorders. *Mov. Disord. Off. J. Mov. Disord. Soc.* **24 Suppl 2**, S691–698 (2009).
306. Kikuchi, A. *et al.* In vivo visualization of alpha-synuclein deposition by carbon-11-labelled 2-[2-(2-dimethylaminothiazol-5-yl)ethenyl]-6-[2-(fluoro)ethoxy]benzoxazole positron emission tomography in multiple system atrophy. *Brain J. Neurol.* **133**, 1772–1778 (2010).
307. Mitsui, J., Matsukawa, T., Yasuda, T., Ishiura, H. & Tsuji, S. Plasma Coenzyme Q10 Levels in Patients With Multiple System Atrophy. *JAMA Neurol.* **73**, 977–980 (2016).
308. Kasai, T. *et al.* Serum Levels of Coenzyme Q10 in Patients with Multiple System Atrophy. *PloS One* **11**, e0147574 (2016).
309. Hu, X., Yang, Y. & Gong, D. Cerebrospinal fluid levels of neurofilament light chain in multiple system atrophy relative to Parkinson's disease: a meta-analysis. *Neurol. Sci. Off. J. Ital. Neurol. Soc. Ital. Soc. Clin. Neurophysiol.* (2016).  
doi:10.1007/s10072-016-2783-7
310. Marques, T. M. *et al.* MicroRNAs in Cerebrospinal Fluid as Potential Biomarkers for Parkinson's Disease and Multiple System Atrophy. *Mol. Neurobiol.* (2016).  
doi:10.1007/s12035-016-0253-0

311. Welter, D. *et al.* The NHGRI GWAS Catalog, a curated resource of SNP-trait associations. *Nucleic Acids Res.* **42**, D1001–1006 (2014).
312. Ferrari, R. *et al.* Frontotemporal dementia and its subtypes: a genome-wide association study. *Lancet Neurol.* **13**, 686–699 (2014).
313. Van Deerlin, V. M. *et al.* Common variants at 7p21 are associated with frontotemporal lobar degeneration with TDP-43 inclusions. *Nat. Genet.* **42**, 234–239 (2010).
314. Moebus, S. *et al.* with Lambert, J. C. *et al.*, European Alzheimer’s Disease Initiative (EADI), Genetic and Environmental Risk in Alzheimer’s Disease, Alzheimer’s Disease Genetic Consortium, and Cohorts for Heart and Aging Research in Genomic Epidemiology. Meta-analysis of 74,046 individuals identifies 11 new susceptibility loci for Alzheimer’s disease. *Nat. Genet.* **45**, 1452–1458 (2013).
315. Lee, S. H., Wray, N. R., Goddard, M. E. & Visscher, P. M. Estimating missing heritability for disease from genome-wide association studies. *Am. J. Hum. Genet.* **88**, 294–305 (2011).
316. Purcell, S. *et al.* PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am. J. Hum. Genet.* **81**, 559–575 (2007).
317. Laurie, C. C. *et al.*, GENEVA Investigators. Quality control and quality assurance in genotypic data for genome-wide association studies. *Genet. Epidemiol.* **34**, 591–602 (2010).
318. Federoff, M. *et al.* Genome-wide estimate of the heritability of Multiple System Atrophy. *Parkinsonism Relat. Disord.* **22**, 35–41 (2016).

319. Howie, B., Fuchsberger, C., Stephens, M., Marchini, J. & Abecasis, G. R. Fast and accurate genotype imputation in genome-wide association studies through pre-phasing. *Nat. Genet.* **44**, 955–959 (2012).
320. Ahmed, Z. *et al.* The neuropathology, pathophysiology and genetics of multiple system atrophy. *Neuropathol. Appl. Neurobiol.* **38**, 4–24 (2012).
321. Bower, J. H., Maraganore, D. M., McDonnell, S. K. & Rocca, W. A. Incidence of progressive supranuclear palsy and multiple system atrophy in Olmsted County, Minnesota, 1976 to 1990. *Neurology* **49**, 1284–1288 (1997).
322. Osaki, Y., Ben-Shlomo, Y., Lees, A. J., Wenning, G. K. & Quinn, N. P. A validation exercise on the new consensus criteria for multiple system atrophy. *Mov. Disord. Off. J. Mov. Disord. Soc.* **24**, 2272–2276 (2009).
323. Sailer, A. *et al.* A genome-wide association study in multiple system atrophy. *Neurology* **87**, 1591–1598 (2016).
324. Potashkin, J. A., Santiago, J. A., Ravina, B. M., Watts, A. & Leontovich, A. A. Biosignatures for Parkinson's disease and atypical parkinsonian disorders patients. *PloS One* **7**, e43595 (2012).
325. Poewe, W. & Wenning, G. The differential diagnosis of Parkinson's disease. *Eur. J. Neurol.* **9 Suppl 3**, 23–30 (2002).
326. Quinn, N. P. How to diagnose multiple system atrophy. *Mov. Disord. Off. J. Mov. Disord. Soc.* **20 Suppl 12**, S5–S10 (2005).
327. Hardy, J. & Singleton, A. Genomewide association studies and human disease. *N. Engl. J. Med.* **360**, 1759–1768 (2009).

328. Pearson, T. A. & Manolio, T. A. How to interpret a genome-wide association study. *JAMA* **299**, 1335–1344 (2008).
329. Keller, M. F. *et al.* Genome-wide analysis of the heritability of amyotrophic lateral sclerosis. *JAMA Neurol.* **71**, 1123–1134 (2014).
330. Al-Chalabi, A. *et al.* An estimate of amyotrophic lateral sclerosis heritability using twin data. *J. Neurol. Neurosurg. Psychiatry* **81**, 1324–1326 (2010).
331. Wirdefeldt, K., Adami, H.-O., Cole, P., Trichopoulos, D. & Mandel, J. Epidemiology and etiology of Parkinson's disease: a review of the evidence. *Eur. J. Epidemiol.* **26 Suppl 1**, S1–58 (2011).
332. Bandrés-Ciga, S. *et al.* Genome-wide assessment of Parkinson's disease in a Southern Spanish population. *Neurobiol. Aging* **45**, 213.e3–9 (2016).
333. Kara, E. *et al.* Assessment of Parkinson's disease risk loci in Greece. *Neurobiol. Aging* **35**, 442.e9–442.e16 (2014).
334. Hernandez, D. G. *et al.* Genome wide assessment of young onset Parkinson's disease from Finland. *PloS One* **7**, e41859 (2012).
335. Yang, J. *et al.* Common SNPs explain a large proportion of the heritability for human height. *Nat. Genet.* **42**, 565–569 (2010).
336. Ku, C. S. *et al.* A new paradigm emerges from the study of de novo mutations in the context of neurodevelopmental disease. *Mol. Psychiatry* **18**, 141–153 (2013).
337. Whiteford, N. *et al.* Swift: primary data analysis for the Illumina Solexa sequencing platform. *Bioinforma. Oxf. Engl.* **25**, 2194–2199 (2009).
338. McCarthy, M. I. *et al.* Genome-wide association studies for complex traits: consensus, uncertainty and challenges. *Nat. Rev. Genet.* **9**, 356–369 (2008).

339. Cheng, K. F., Lee, J. Y., Zheng, W. & Li, C. A powerful association test of multiple genetic variants using a random-effects model. *Stat. Med.* **33**, 1816–1827 (2014).
340. Lee, S., Wu, M. C. & Lin, X. Optimal tests for rare variant effects in sequencing association studies. *Biostat. Oxf. Engl.* **13**, 762–775 (2012).
341. Li, B. & Leal, S. M. Methods for detecting associations with rare variants for common diseases: application to analysis of sequence data. *Am. J. Hum. Genet.* **83**, 311–321 (2008).
342. Lesage, S. *et al.*, French Parkinson's Disease Genetics Study (PDG), International Parkinson's Disease Genomics Consortium (IPDGC), and International Parkinson's Disease Genomics Consortium IPDGC. Loss of VPS13C Function in Autosomal-Recessive Parkinsonism Causes Mitochondrial Dysfunction and Increases PINK1/Parkin-Dependent Mitophagy. *Am. J. Hum. Genet.* **98**, 500–513 (2016).
343. Wang, L. *et al.* Association of four new candidate genetic variants with Parkinson's disease in a Han Chinese population. *Am. J. Med. Genet. Part B Neuropsychiatr. Genet. Off. Publ. Int. Soc. Psychiatr. Genet.* **171**, 342–347 (2016).
344. Vennemann, A. & Hofmann, T. G. SUMO regulates proteasome-dependent degradation of FLASH/Casp8AP2. *Cell Cycle Georget. Tex* **12**, 1914–1921 (2013).
345. Tsou, J.-H. *et al.* Important Roles of Ring Finger Protein 112 in Embryonic Vascular Development and Brain Functions. *Mol. Neurobiol.* (2016).  
doi:10.1007/s12035-016-9812-7
346. Feng, S., Liu, D., Zhan, X., Wing, M. K. & Abecasis, G. R. RAREMETAL: fast and powerful meta-analysis for rare variants. *Bioinforma. Oxf. Engl.* **30**, 2828–2829 (2014).

347. Lachance, J. & Tishkoff, S. A. SNP ascertainment bias in population genetic analyses: why it is important, and how to correct it. *BioEssays News Rev. Mol. Cell. Dev. Biol.* **35**, 780–786 (2013).

## 8 Appendix

### 8.1.1 Transcripts

#### 8.1.1.1 MSA local pipeline: WES variant filtering approach

#CHROM	POSITION	REF	ALT	GENE	ID	EFFECT	Transcript
8	16859372	AGTCTGTG		FGF20	.	frameshift insertion	FGF20:NM_019851:exon1:c.169_170insCACAGACT:p.L57fs,
12	40749989	C		LRRK2	.	frameshift insertion	LRRK2:NM_198578:exon46:c.6844dupC:p.K2281fs,

Table 38: PD genes in MSA VCF that did not confirm with Sanger sequencing

#CHROM	POS	REF	ALT	ID	GENE NAME	EFFECT	TRANSCRIPT
1	6727802	TTC	T	.	DNAJC11	Frameshift	NM_018198.3 4 1
1	12921594	C	G	rs201429745	PRAMEF2	Stop_gained	NM_023014.1 4 1 tCa tGa 5462* 474
2	97749524	C	A	rs187104986	FAHD2B	Stop_gained	NM_199336.1 8 1 Gaa Taa E305* 314
2	113416606	CAG	C	.	SLC20A1	Frameshift	NM_005415.4 7 1
5	130815368	C	CT	.	RAPGEF6	Frameshift	NM_001164389.1 16 1 NM_001164388.1 NM_001164387.1 16 1 NM_016340.5 16 1 NM_001164386.1 16 1 NM_001164390.1 16 1
5	145647319	T	TA	.	RBM27	Frameshift	NM_018989.1 15 1
6	108985176	T	TG	rs34133353	FOXO3	Frameshift	NM_001455.3 2 1 NM_201559.2 3 1
7	87074281	G	GA	.	ABC84	Frameshift	NM_018850.2 10 1
9	35059646	A	AT	.	VCP	Frameshift	NM_007126.3 14 1
10	78729785	C	CT	.	KCNMA1	Frameshift	NM_001271518.1 19 1
15	41862897	G	GGAGA	.	TYRO3	Splice site donor	NM_006293.3 12 1
15	93540315	G	GA	.	CHD2	Frameshift	NM_001271.3
12	76424933	GC	G	.	PHLDA1	Frameshift	NM_007350.3 1 1
1	46184896	AAC	A	rs144663569	IIP	Frameshift	NM_001145349.1 6 1
2	232577559	T	C	.	PTMA	Stop_lost	NM_002823.4 5 1 Tag Cag *112Q 111
X	55172630	G	A	rs113263757	FAM104B	Stop_gained	NR_030722.1 3 1 Cag Tag Q76* 112 NM_001166702.1 3 1
11	6567913	C	CT	.	DNHD1	Frameshift	NM_144666.2 21 1
2	96780997	CAG	C	.	ADRA2B	Frameshift	NM_000682.5 1 1

Table 39: Variants predicted very damaging in MSA VCF not confirmed with Sanger sequencing

#### 8.1.1.2 MSA local pipeline: WES gene filtering approach

#CHROM	POSITION	REF	ALT	GENE NAME	ID	Effect	Transcript
chr22	18121597	C	T	BCCL2L13	.	nonynonymous SNV	BCCL2L13:NM_001270727:exon1:c.C98T;p.P31L;BCCL2L13:NM_001270726:exon1:c.C98T;p.P31L
chr22	18121597	G	T	BCCL2L13	.	nonynonymous SNV	BCCL2L13:NM_001270727:exon1:c.G128T;p.G43V;BCCL2L13:NM_001270726:exon1:c.G128T;p.G43V
chr6	96573431	T	C	CASPA2	.	nonynonymous SNV	CASPA2:NM_001137668:exon7:c.T2003C;p.M668T;CASPA2:NM_001137667:exon7:c.T2003C;p.M668T;CASPA2:NM_012115:exon7:c.T2003C;p.M668T
chr6	96577755	A	G	CASPA2	.	nonynonymous SNV	CASPA2:NM_001137668:exon8:c.A4766G;p.N1589S;CASPA2:NM_001137667:exon8:c.A4766G;p.N1589S;CASPA2:NM_012115:exon8:c.A4766G;p.N1589S
chr6	96572365	T	G	CASPA2	.	nonynonymous SNV	CASPA2:NM_001137668:exon7:c.T937G;p.Y131D;CASPA2:NM_001137667:exon7:c.T937G;p.Y131D;CASPA2:NM_012115:exon7:c.T937G;p.Y131D
chr11	11373751	G	A	CSNCA3	.	stopgain SNV	CSNCA3:NM_001256686:exon1:c.C97T;p.R396K
chr11	11373910	C	A	CSNCA3	.	nonynonymous SNV	CSNCA3:NM_001256686:exon1:c.G57T;p.D259Y
chr2	53322243	G	A	DCPIA	.	nonynonymous SNV	DCPIA:NM_001290205:exon10:c.C1287T;p.A43V;DCPIA:NM_0018408:exon9:c.C158T;p.A50V;DCPIA:NM_001290204:exon8:c.C1394T;p.A465V;DCPIA:NM_001290207:exon7:c.C1031T;p.A344V;DCPIA:NM_001290206:exon6:c.C1145T;p.A382V
chr2	53373290	G	A	DCPIA	.	nonynonymous SNV	DCPIA:NM_0018408:exon2:c.C185T;p.S63E;DCPIA:NM_001290204:exon3:c.C185T;p.S63E
chr7	1786179	T	G	ELFN1	.	nonynonymous SNV	ELFN1:NM_001128636:exon2:c.T2399G;p.L1808R
chr7	1785179	G	A	ELFN1	.	nonynonymous SNV	ELFN1:NM_001128636:exon2:c.G947A;p.R316H
chr7	159434680	G	A	GIMAP1-GIMAP5	.	nonynonymous SNV	GIMAP1-GIMAP5:NM_001199577:exon6:c.G480A;p.A164T
chr15	23892819	C	T	MAGEL2	.	nonynonymous SNV	MAGEL2:NM_010966:exon1:c.G71A;p.S24N
chr15	23898934	C	A	MAGEL2	.	nonynonymous SNV	MAGEL2:NM_010966:exon1:c.G329G;p.A106S
chr19	1527953	C	T	PLK3	.	nonynonymous SNV	PLK3:NM_001249079:exon7:c.C237T;p.TK
chr19	1528003	G	C	PLK3	.	nonynonymous SNV	PLK3:NM_001249079:exon7:c.G71C;p.R240E
chr8	145741895	C	T	RECQL4	.	nonynonymous SNV	RECQL4:NM_004260:exon5:c.G888A;p.G201E
chr8	145739330	C	T	RECQL4	.	nonynonymous SNV	RECQL4:NM_004260:exon12:c.G3940A;p.M680I
chr8	145740372	C	G	RECQL4	.	nonynonymous SNV	RECQL4:NM_004260:exon9:c.G1568C;p.S532E
chr1	182496832	G	C	RGSL1	.	nonynonymous SNV	RGSL1:NM_001137669:exon11:c.G289C;p.E944Q
chr1	182501804	G	C	RGSL1	.	nonynonymous SNV	RGSL1:NM_001137669:exon14:c.G2372C;p.G791A
chr1	18244327	G	T	RGSL1	.	nonynonymous SNV	RGSL1:NM_001137669:exon16:c.G1261T;p.L427E
chr1	182522662	G	T	RGSL1	.	nonynonymous SNV	RGSL1:NM_001137669:exon19:c.G3108C;p.S109T
chr17	19181158	C	A	RNF112	.	nonynonymous SNV	RNF112:NM_007148:exon10:c.C1084A;p.P362I
chr19	5395966	C	T	ZNF761	.	nonynonymous SNV	ZNF761:NM_001289951:exon6:c.C1705T;p.R596C;ZNF761:NM_001008401:exon7:c.C1705T;p.R596C;ZNF761:NM_001289952:exon6:c.C1705T;p.R596C
chr19	5395985	G	A	ZNF761	.	nonynonymous SNV	ZNF761:NM_001289951:exon6:c.GK24A;p.C275Y;ZNF761:NM_001008401:exon7:c.GK24A;p.C275Y;ZNF761:NM_001289952:exon6:c.GK24A;p.C275Y
chr7	6661799	C	G	ZNF853	.	nonynonymous SNV	ZNF853:NM_017560:exon3:c.C1277G;p.L393V

Table 40: All variants in MSA VCF confirmed with Sanger sequencing

#CHROM	POS	REF	ALT	GENE NAME	ID	TRANSCRIPT ID
chr22	18121566	G	A	BCL2L13	.	nonsynonymous SNV BCL2L13:NM_001270727:exon1:c.G107A:p.G36E,BCL2L13:NM_001270726:exon1:c.G107A:p.G36E,
chr3	53381504	G	A	DCP1A	.	nonsynonymous SNV DCP1A:NM_018403:exon1:c.C41T:p.A14V,DCP1A:NM_001290204:exon1:c.C41T:p.A14V,
chr8	143747275	C	G	JRK	.	nonsynonymous SNV JRK:NM_003724:exon2:c.G203C:p.S68T,JRK:NM_001279352:exon2:c.G203C:p.S68T,JRK:NM_001077527:exon2:c.G203C:p.S68T,
chr8	143745974	G	T	JRK	.	nonsynonymous SNV JRK:NM_003724:exon3:c.C1502A:p.A501E,JRK:NM_001279352:exon3:c.C1502A:p.A501E,JRK:NM_001077527:exon3:c.C1502A:p.A501E,
chr1	17721517	G	A	PADI6	.	nonsynonymous SNV PADI6:NM_207421:exon13:c.G1408A:p.E470K,
chr1	182441573	T	A	RGS1	.	nonsynonymous SNV RGS1:NM_001137669:exon5:c.T344A:p.L115Q,
chr7	6661490	C	T	ZNF853	.	stopgain SNV ZNF853:NM_017560:exon3:c.C868T:p.Q290X,
chr7	6662446	C	A	ZNF853	.	nonsynonymous SNV ZNF853:NM_017560:exon3:c.C1824A:p.H608Q,
chr9	115759981	G	A	ZNF883	.	nonsynonymous SNV ZNF883:NM_001101338:exon5:c.C559T:p.R187W,

**Table 41: All variants that did not confirm with Sanger sequencing**

### 8.1.1.3 MSA Googlegenome pipeline: Gene burden and single variant analyses

CHR	POSITION	ID	REF	ALT	GENE	EFFECT	TRANSCRIPT
12	40757328	rs34778348	G	C	LRRK2	Non-synonymous	NM_198578:(GGA/Gly/G->CGA/Arg/R:Base7154/7584:Codon2385/2528:Exon48/51)
1	75684185	rs200948281	G	C	SLC44A5	Non-synonymous	NM_001130058:(CGA/Arg/R->GGA/Gly/G:Base1520/2154:Codon507/718:Exon17/24)/NM_152697:-(CGA/Arg/R->GGA/Gly/G:Base1520/2160:Codon507/720:Exon17/24)
12	75874811	rs76259505	A	C	GLIPR1	Non-synonymous	NM_006851:(ACA/Thr/T->CCA/Pro/P:Base152/801:Codon51/267:Exon1/6)
6	90577876	rs150022229	G	T	CASP8AP2	Non-synonymous	NM_012115(GTT/Val/V->TTT/Phe/F:Base4868/5901:Codon1623/1967:Exon8/10)/NM_001137668(GTT/Val/V->TTT/Phe/F:Base4868/5901:Codon1623/1967:Exon8/10)

**Table 42: All variants investigated in the “in-depth gene” analysis.**

### 8.1.1.4 Greek Rapsani PD WES candidates

#CHROM	POS	REF	ALT	GENE	ID	Effect	TRANSCRIPT
chr12	88481560	G	T	CEP290	.	nonsynonymous SNV	NM_025114:exon32:c.C4191A:p.N1397K
chr7	151945071	-	T	KMT2C	.	stopgain SNV	NM_170606:exon14:c.2447dupA;p.Y816_I817delinsX
chr11	66335541	A	C	CTSF	rs200958879	nonsynonymous SNV	NM_003793:exon2:c.T226G:p.S76A

**Table 43: Variants checked with Sanger sequencing in Greek PD cohort from WES data**

Chr	POS	ID	REF	ALT	Gene	Effect	Transcript
chr16	25251790	-	G	C	ZKSCAN2	NS SNV	NM_001012981:exon7:c.G251G:p.L751V
chr4	266147	rs202099832	T	C	ZNF732	NS SNV	NM_001137608:exon3:c.A496G:p.K166E
chr9	14859241	-	C	T	FREM1	NS SNV	NM_144966:exon5:c.G371A:p.G181R
chr11	125322281	-	C	T	FEZ1	NS SNV	NM_005103:exon8:c.G1045A:p.E349K
chr11	95826086	rs191030213	G	A	MAML2	NS SNV	NM_032427:exon2:c.C119T:p.P370L
chr2	236626225	rs201400274	G	A	AGAP1	NS SNV	NM_001097331:exon3:c.G247A:p.G83S,NM_001244888:exon3:c.G247A:p.G83S
chr12	581120546	-	T	C	AGAP2	NS SNV	NM_014770:exon18:c.A1300G:p.Q767K,NM_001122772:exon19:c.A3368G:p.Q1123R
chr15	75042189	-	A	G	CYP1A2	NS SNV	NM_000761:exon2:c.A119G:p.X37R
chr12	22199386	-	T	C	CMA5	NS SNV	NM_018668:exon1:c.T149C:p.L50P
chr17	1386937	rs146471734	C	T	MYO1C	NS SNV	NM_001080950:exon3:c.G257A:p.R86H,NM_001080779:exon3:c.G314A:p.R105H,NM_003375:exon3:c.G209A:p.R70H
chr14	101005265	-	C	T	BEGAIN	NS SNV	NM_001159531:exon6:c.G823A:p.A275T,NM_020836:exon7:c.G823A:p.A275T
chr19	54409660	-	C	A	PRKCG	NS SNV	NM_002739:exon17:c.C1854A:p.D618E
chr12	57351191	rs199814447	C	T	RDH16	NS SNV	NM_003708:exon1:c.G56A:p.R19Q
chr4	153896551	rs147950044	C	T	FHDC1	NS SNV	NM_033993:exon11:c.C2108T:p.P709L
chr12	118693338	rs78662524	G	T	TAOK3	NS SNV	NM_016281:exon3:c.C35A:p.A12D
chr7	6217468	-	G	C	CYTH3	NS SNV	NM_004227:exon5:c.C354G:p.D118E
chr6	158517100	-	G	C	SYNJ2	NS SNV	NM_003898:exon27:c.G4195C:p.A1399P,NM_001178088:exon26:c.G3484C:p.A1162P
chr16	15703534	rs200431093	A	G	KIAA0430	NS SNV	NM_001184999:exon20:c.T7391C:p.V1264A,NM_014647:exon20:c.T3800C:p.V1267A,NM_001184998:exon20:c.T3800C:p.V1267A
chr16	30774746	-	A	T	RNF40	NS SNV	NM_014771:exon4:c.A308T:p.L103V,NM_001286572:exon4:c.A308T:p.L103V,NM_001207033:exon4:c.A308T:p.L103V,NM_001207034:exon4:c.A308T:p.L103V
chr19	44738864	-	A	T	ZNF227	NS SNV	NM_001289187:exon6:c.A128T:p.H43L,NM_001289171:exon6:c.A447p.H33L,NM_182490:exon6:c.A281T:p.H94,NM_001289168:exon5:c.A128T:p.H43L
chr15	91454462	rs200259502	C	T	MAN2A2	NS SNV	NM_006122:exon12:c.C1937T:p.S646L
chr16	18865003	rs184038326	C	T	SMG1	NS SNV	NM_015052:exon31:c.G4670A:p.G1357D
chr3	108366806	-	G	C	DZIP3	NS SNV	NM_014648:exon16:c.G1809C:p.E603D
chr5	148427540	-	G	T	SH3TC2	NS SNV	NM_024577:exon3:c.C164A:p.S55
chr17	74037057	-	G	C	SRP68	NS SNV	NM_001260503:exon7:c.C510G:p.D170E,NM_014230:exon14:c.C1527G:p.D509E,NM_001260502:exon13:c.C1413G:p.D471E

**Table 44: Variants checked with Sanger sequencing in Greek PD individual families from WES data**



### 8.1.1.5 Greek Rapsani PD WGS candidates

CHR	POSITION	ID	REF	ALT	GENE	EFFECT	Transcript
chr13	78272267	.	-	GG	SLAIN1	Frameshift Insertion	NM_001242868:exon1:c.219_220insGG:p.A73fs
chr16	70896016	.	A	-	HYDIN1	Frameshift Deletion	NM_001270974:exon69:c.11712delT:p.I3904fs
chr16	70972620	.	G	C	HYDIN1	Non-synonymous coding	NM_001270974:exon44:c.C6892G:p.R2298G
chr9	70912543	.	A	T	CBWD3	Non-synonymous coding	NM_201453:exon12:Amino Acid:288
chr13	72440658	.	TGCCGCT	.	DACH1	Codon Deletion	NM_080759:exon1:c.244_249del:p.82_83del, NM_004392:exon1:c.244_249del:p.82_83del, NM_080760:exon1:c.244_249del:p.82_83del
chr14	67940153	.	T	C	TMEM229B	Non-synonymous coding	NM_182526:exon3:c.A488G:p.H163R
chr14	68038891	.	C	G	PLEKHH1	Non-synonymous coding	NM_020715:exon11:c.C1625G:p.A542G

**Table 45: Variants checked with Sanger sequencing in Greek PD cohort from WGS data**

## 8.1.2 Primer sequences

### 8.1.2.1 MSA WES variant filtering approach candidates

#CHROM	POS	REF	ALT	GENE NAME	ID	EFFECT	PRIMER SEQUENCE (F,R)
8	16859372	AGTCTGTG	.	FGF20	.	Frameshift Insertion	CAAAGAGGGAGGTGCAAGG, GCCATTTCACTGCAAGTCC
12	40749989	C	.	LRRK2	.	Frameshift Insertion	GGGATTTCACCATGTGTAGCC, TGGCAATAGGAGAATGAACG

**Table 46: PD genes in MSA VCF that did not confirm with Sanger sequencing**

#CHROM	POS	REF	ALT	GENE NAME	ID	EFFECT	PRIMER SEQUENCE(F,R)
1	6727802	TTC	T	DNAJC11	.	Frameshift	GTTAGGTGGCTGTCTCTGC, TCCTGAGGAAAGGCTGTGG
1	12921594	C	G	PRAMEF2	rs201429745	Stop_gained	CGGGCTGAGCTGATGTGTA, CCTTGAAACAGGTGCCATTT
2	97749524	C	A	FAHD2B	rs187104986	Stop_gained	GAGGCTGCAAACCTTAGC, GTGAGTGACAAGGGCTGTCC
2	113416606	CAG	C	SLC20A1	.	Frameshift	AGTATGGCCACAACCAACC, GAGTCTCCCATGCAATCTCC
5	130815368	C	CT	RAPGEF6	.	Frameshift	CAAAGCAAACTTACCCAAGC, TCAAACCTGGTGTGTCAAGG
5	145647319	T	TA	RBM27	.	Frameshift	TTGAGCCCAAGAGATTGAGG, TCCAAGTTTACATTAGAGAAGAAGC
6	108985176	T	TG	FOXO3	rs34133353	Frameshift	TTTTCTCCTGCAGAACTCC, GAATTCTGGGCAGACACAGC
7	87074281	G	GA	ABC4	.	Frameshift	TACATAGGGCCAAAATCTGC, GAGAAACGCTTTTGCTCTGC
9	35059646	A	AT	VCP	.	Frameshift	TAGCCAGGAACTCCAAGTCC, TCTCAGTATGTTGCCCATGC
10	78729785	C	CT	KCNMA1	.	Frameshift	TTTTAGCCTGAGCAGAACTGG, CTGGGCAAGCAAGTAACC
15	41862897	G	GGAGA	TYRO3	.	Splice site donor	ATCCCTCTTCCCAACAACC, TCCACTTGCTAATGCCTTCC
15	93540315	G	GA	CHD2	.	Frameshift	GAAAAGGACCAGGGAAAAGG, GAAACAGTGCCACAATCTGC
12	76424933	GC	G	PHLDA1	.	Frameshift	TGGATGTGGATGTGGATGC, CTCTTATGCTGGGAGGATGC
1	46184896	AAC	A	IPP	rs144663569	Frameshift	GAGTAGAAGGATGGTTACCAGAGG, TGTGCCTCTGTTTCATAAGTGG
2	232577559	T	C	PTMA	.	Stop_lost	CTTGACATCCACTACATGTTCC, ACAGAGAACGCTCCGAAGG
X	55172630	G	A	FAM104B	rs113263757	Stop_gained	TTTCATCATGTTGGAGCTTGC, CAAGGCCTGTGTTGAATCC
11	6567913	C	CT	DNHD1	.	Frameshift	AATGTGACCAAGGAGGAACC, CTGGCTCAACTGTTCCAAGG
2	96780997	CAG	C	ADRA2B	.	Frameshift	GGTCCTCAAGAAAGGGAAGC, AGCATCGGATCTTTCTTTC

**Table 47: Variants predicted very damaging in MSA VCF not confirmed with Sanger sequencing**

### 8.1.2.2 MSA WES gene filtering approach candidates

#CHROM	POSITION	REF	ALT	GENE NAME	ID	EFFECT	PRIMER SEQUENCE (F, R)
chr22	18121557	C	T	<i>BCL2L13</i>	.	nonsynonymous SNV	TATTTGACCTGCTCCCTTGG, TCTCTGGGCACTTTCCTACC
chr22	18121587	G	T	<i>BCL2L13</i>	.	nonsynonymous SNV	TATTTGACCTGCTCCCTTGG, TCTCTGGGCACTTTCCTACC
chr6	90573431	T	C	<i>CASP8AP2</i>	.	nonsynonymous SNV	CAACGTTATTGGAACCAAAAGG, TTGGCAAAATATCAGCTTCC
chr6	90577775	A	G	<i>CASP8AP2</i>	.	nonsynonymous SNV	AAATACGACGTGCAACACCA, GGAAGATGTTCCCCAACAGA
chr6	90572365	T	G	<i>CASP8AP2</i>	.	nonsynonymous SNV	TATCAGGTTGGCGAGGGTAG, GAGGAAGTGATGCTCGTTTACAG
chr11	11373751	G	A	<i>CSNK2A3</i>	.	stopgain SNV	ATCTTTTCGGAAGGAGCCATT, GGCACCTGAAGAAATCCCTGA
chr11	11373910	C	A	<i>CSNK2A3</i>	.	nonsynonymous SNV	ATCTTTTCGGAAGGAGCCATT, GGCACCTGAAGAAATCCCTGA
chr3	53322243	G	A	<i>DCP1A</i>	.	nonsynonymous SNV	ATCCTGGAGCTGAGACTTGC, ACTGTGGTGGTTGGACTTGG
chr3	53376290	G	A	<i>DCP1A</i>	.	nonsynonymous SNV	CTTGGTGGCCTCAGTAAGGA, GAAATGGTTCATGGAGCTGAA
chr7	1786631	T	G	<i>ELFN1</i>	.	nonsynonymous SNV	GGCCAAGTACATCGAGAAGG, TGTCCCTCTGTCTGTCC
chr7	1785179	G	A	<i>ELFN1</i>	.	nonsynonymous SNV	GTCAGTCTGCACCGAGGACT, GTTGAAATGCTCCAGGGTGT
chr7	150434680	G	A	<i>GIMAP1-GIMAP5</i>	.	nonsynonymous SNV	ATCAGTTTCCAGCCAACACC, ACCCAGAAGTGAGGGGATCT
chr15	23892819	C	T	<i>MAGEL2</i>	.	nonsynonymous SNV	GCCACGTAGGCATTCTCTTC, GAGTCGGAGGCTTACCCATC
chr15	23889634	C	A	<i>MAGEL2</i>	.	nonsynonymous SNV	CCAAAAACCTCAGGACAAGC, TGTCTCCCTTGGATGAGAGG
chr19	1527955	C	T	<i>PLK5</i>	.	nonsynonymous SNV	GTGGTGTGCACCTGTAGTCC, GAGTGAACCCCTTGGATGG
chr19	1528003	G	C	<i>PLK5</i>	.	nonsynonymous SNV	GGATGTGGGAGATGAGGAA, CCCACCTGTGTGAAGAAGT
chr8	145741895	C	T	<i>RECQL4</i>	.	nonsynonymous SNV	GCCTCATCTAAGGCATCCAC, ACGGATGCTGACTTCTTGA
chr8	145739330	C	T	<i>RECQL4</i>	.	nonsynonymous SNV	GGGATGGGACCATGTGTG, TACACGTGCTGATGCTGACA
chr8	145740372	C	G	<i>RECQL4</i>	.	nonsynonymous SNV	GCTCCCATTTACCTCTCTCC, AAGTGTCTGGTCTTGGCTGT
chr1	182496832	G	C	<i>RGSL1</i>	.	nonsynonymous SNV	GTCATATGCACCTCCCATTT, TGTGGCTGCTCTGTACT
chr1	182501804	G	C	<i>RGSL1</i>	.	nonsynonymous SNV	CTTTACGGAGGTGTTGTGAGG, GCTTCTTATGTGGGCTGACC
chr1	182443527	G	T	<i>RGSL1</i>	.	nonsynonymous SNV	GGACTGTACATCCCATTTGTC, CCTTGCAAATCTCCTTCTGG
chr1	182522662	G	C	<i>RGSL1</i>	.	nonsynonymous SNV	GGAGTTGTGCGAGGTGTTTT, CGGGGTTACAAAAACCAAGA
chr17	19318158	C	A	<i>RNF112</i>	.	nonsynonymous SNV	AGGACCCCTTCTTTTACAGC, CTCCAGGGAGTTTTCACACC
chr19	53959466	C	T	<i>ZNF761</i>	.	nonsynonymous SNV	GGGTAACTGCTGATTGAAGG, GGCAGAAAGTCAATCTTACACG
chr19	53958585	G	A	<i>ZNF761</i>	.	nonsynonymous SNV	CTTGGATCAAGCTTTCATTTC, CAAAATTGATCTGAAACTGTAAGC
chr7	6661799	C	G	<i>ZNF853</i>	.	nonsynonymous SNV	CAAGATGGGCAACAACAGC, CTGCACCACCACGTAGCC

Table 48: All variants in MSA VCF confirmed with Sanger sequencing

#CHROM	POS	REF	ALT	GENE NAME	ID	EFFECT	PRIMER SEQUENCE (F,R)
chr22	18121566	G	A	<i>BCL2L13</i>	.	nonsynonymous SNV	TATTTGACCTGCTCCCTTGG, TCTCTGGGCACTTTCCTACC
chr3	53381504	G	A	<i>DCP1A</i>	.	nonsynonymous SNV	ACCAGAATGGCTGACGTACC, AGGCTCTGGCTAGGCTCTTG
chr8	143747275	C	G	<i>JRK</i>	.	nonsynonymous SNV	AGAGGGAAGTAGTGGCAGCA, AAGTCCTTGGCCTTCTCGAT
chr8	143745974	G	T	<i>JRK</i>	.	nonsynonymous SNV	CTGGAGCTTGTGAAGGAAGG, GTTGTACCTGCTGTGGATG
chr1	17721517	G	A	<i>PADI6</i>	.	nonsynonymous SNV	TTTACACAACGGTCAGAGG, GGTGAGATGTCCTTCTTCCA
chr1	182441573	T	A	<i>RGSL1</i>	.	nonsynonymous SNV	TTCCCATGCAATTAAGCCTCT, GGTGTGCACATGCTCTCTG
chr7	6661490	C	T	<i>ZNF853</i>	.	stopgain SNV	CAAGATGGGCAACAACAGC, CTGCACCACCACGTAGCC
chr7	6662446	C	A	<i>ZNF853</i>	.	nonsynonymous SNV	ACCCTACGCTGCTCTACT, GCGGTGAAATCTGTGAACCT
chr9	115759981	G	A	<i>ZNF883</i>	.	nonsynonymous SNV	AATGAAGGCTGGGGTATGG, CTTGTGTGCATCCCTGACC

Table 49: All variants that did not confirm with Sanger sequencing

### 8.1.2.3 MSA Googlegenome pipeline: Gene burden and single variant analyses

CHR	POSITION	ID	REF	ALT	GENE	EFFECT	PRIMER SEQUENCE (F, R)
12	40757328	rs34778348	G	C	LRK2	Non-synonymous	TGCAGCTTTCAGTGATTCCA, TTTTCATTGAGAACATTTCTTG
1	75684185	rs200948281	G	C	SLC44A5	Non-synonymous	GGGGCTCTGTGTAACCTTGC, AGACAAAACCAATGGGCATC
12	75874811	rs76259505	A	C	GLIPR1	Non-synonymous	CATGCGTGTACACTTGTCTA, AAAATGTGCTGTGGCACTTG
6	90577876	rs150022229	G	T	CASP8AP2	Non-synonymous	AAATACGACGTGCAACACCA, GGAAGATGTTCCCCAACAGA

### 8.1.2.4 Greek Rapsani PD WES candidates

#CHROM	POS	REF	ALT	GENE	ID	Effect	Primer Sequence (F, R)
chr12	88481560	G	T	CEP290	.	nonsynonymous SNV	TTAAATCTTAGGGGCTCACG, TGCTTATAGTCAGTGTTCATAAACG
chr7	151945071	-	T	KMT2C	.	stopgain SNV	GTTTGGACCGAGGTCTACCA, GTCTTCTCCCCAACACCTT
chr11	66335541	A	C	CTSF	rs200958879	nonsynonymous SNV	TTGACCTAACATTGGGATGG, ACACCATGGGGTCGTTGC

**Table 50: Variants checked with Sanger sequencing in Greek PD cohort from WES data**

Chr	POS	ID	REF	ALT	Gene	Effect	Primer Sequence (F, R)
chr16	25251790	.	G	C	ZKSCAN2	NS SNV	TTGGTGAGGTAAATAGGAAAGAGC, TCTGCCTCATCTGTGATTGC
chr4	266147	rs202099832	T	C	ZNF732	NS SNV	GGCATGAAGAGAACAGTGAGG, TGTGGTGCATGCTAGAGG
chr9	14859241	.	C	T	FREM1	NS SNV	CATCATGGTGAAGGTGAGC, GAAGAAGTGGCAGGAAAAGG
chr11	125322281	.	C	T	FEZ1	NS SNV	TACAGCCAGACCCAGAGACC, TGGGATGGAGAGAAATGTGG
chr11	95826086	rs191030213	G	A	MAML2	NS SNV	CAAATGGACCTGGTGATGG, GGCCTCCAAATAAACATGG
chr2	236626225	rs201400274	G	A	AGAP1	NS SNV	AGCTTGTTTTGGCCATCC, TGGTGGTGGTAATCTGAACG
chr12	58120546	.	T	C	AGAP2	NS SNV	GCAACTATACCAAGCAACC, TACCGTTCTCTGCTTTTGG
chr15	75042189	.	A	G	CYP1A2	NS SNV	TCTCCATCTCAACCCTCAGC, TAGCAGGAGGATGAGGAAAGC
chr12	22199386	.	T	C	CMAS	NS SNV	TGAGGTGGTGAGGACTAGC, CAGTACGGAAACAAGGAAGG
chr17	1386937	rs146471734	C	T	MYO1C	NS SNV	GTTCGCACATTCTCTCTGG, GACGGAATTCCTCATGTTGG
chr14	101005265	.	C	T	BEGAIN	NS SNV	CGGCTGTTCAGGTAGATGG, GGTTCATGGCTAACATCAGTCC
chr19	54409660	.	C	A	PRKCG	NS SNV	CTCTTGGTTCTCCATCTGC, TCTGCACCTCCTTTTGTGG
chr12	57351191	rs199814447	C	T	RDH16	NS SNV	CTCATCAGGGCAAACTCTCC, CATCAGGGAGGAAAGTGAGG
chr4	153896551	rs147950044	C	T	FHDC1	NS SNV	AGCCCAATAAGTTCACAGC, CTACCCACAGATCCCAAGC
chr12	118693338	rs78662524	G	T	TAOK3	NS SNV	TAAGGCTGGTCTCGAACTCC, TGCCTAGGCTGGTCTTCC
chr7	6217468	.	G	C	CYTH3	NS SNV	TAATTGAGACGGGCATGTGG, CACCACCCACAAAGTTTCC
chr6	158517100	.	G	C	SYNJ2	NS SNV	TTGTCCCACTTTTCTCAGC, TGAGTCCAAAGGGTCTGAGC
chr16	15703534	rs200431093	A	G	KIAA0430	NS SNV	AGGAATATGGCTTCTCTGACTGG, AGGTTGTGTGTGTTTGGAAAGG
chr16	30774746	.	A	T	RNF40	NS SNV	TCATCGTCAATCGCTACTGG, ACTGCCAAACATGAAATCC
chr19	44738864	.	A	T	ZNF227	NS SNV	ACCCCACTCTTCGTTTACACC, AACTCTTCCCTGAAGACACC
chr15	91454462	rs200259502	C	T	MAN2A2	NS SNV	CTCTTTTCTCTCCCATCC, TGCCTGTCTCTGGACAACC
chr16	18865003	rs184038326	C	T	SMG1	NS SNV	GAGGGGCAGGATCTTATTGC, GTAGGCCAGTCAACACATGC
chr3	108366806	.	G	C	DZIP3	NS SNV	TCTCTGGGGCAATATCAAGG, GGCAGGAAAGATTATGGAAGG
chr5	148427540	.	G	T	SH3TC2	NS SNV	CTCAGCATTCACCTGTTTCC, ACACCAAGTTGAGGGTTACAGG
chr17	74037057	.	G	C	SRP68	NS SNV	ATCAATCCAGATCCAATGC, AAAATGGGAGTGCAAACTGC

**Table 51: Variants checked with Sanger sequencing in Greek PD individual families from WES data**

### 8.1.2.5 Greek Rapsani PD WGS candidates

CHR	POSITION	ID	REF	ALT	GENE	EFFECT	PRIMER SEQUENCE (F, R)
chr13	78272267	.	-	GG	SLAIN1	Frameshift Insertion	CAAGCTGGAGAAGCAGAACG, AGGCAGTACCAGGTGTAGTCG
chr16	70896016	.	A	-	HYDIN1	Frameshift Deletion	GAAAGTAAGCAGGGGGAAGC, TGCAGAATATCACCCACTCC
chr16	70972620	.	G	C	HYDIN1	Non-synonymous coding	CTGCTAGCGAGACTGAGAAGG, TGGGGAGCAGTCACTATGC
chr9	70912543	.	A	T	CBWD3	Non-synonymous coding	CCCTTTTCATAGGGTTTCTG, CAGCCTTATGACCTCCATGC
chr13	72440658	.	TGCCGCC	T	DACHI	Codon Deletion	ATCTCCACGTCTGCTTCCTC, GGTGAGTACACGGGTTTCC
chr14	67940153	.	T	C	TMEM229B	Non-synonymous coding	CAGTTCATCATCCGCAACAC, TGGTACCAACCCCTGAACTG
chr14	68038891	.	C	G	PLEKHH1	Non-synonymous coding	GGGCCATCAAGAGAGGTACA, CTTCCTTGTTGGGTTCTGTGT

**Table 52: Variants checked with Sanger sequencing in Greek PD cohort from WGS data**

### 8.1.3 Cyclers programs

Temperature (°C)	Time (min)	Number of cycles
95	15	X 1
94	0.5	X 2
70	0.5	
72	2	X 3
94	0.5	
68	0.5	X 4
72	2	
94	0.5	X 5
66	0.5	
72	2	X 6
94	0.5	
64	0.5	X 7
72	2	
94	0.5	X 8
62	0.5	
72	0.75	X 7
94	0.5	
60	0.5	X 8
72	2	
94	0.5	X 7
58	0.5	
72	0.75	X 8
94	0.5	
60	0.5	X 5
72	2	
94	0.5	X 1
56	0.5	
72	2	
72	10	

Table 53: 72 touchdown 56 PCR cycler conditions used for all primers. Total time: 2h, 45min

Temperature (°C)	Time (min)	Number of cycles
94	1	x 25
94	0.5	
50	0.25	
60	4	
4	hold	

Table 54: Sequencing cycler conditions used for all primers. Total time: 2h, 22min

### 8.1.4 PCR master mixes

Reagent	Volume per reaction (ul)
Roche Mastermix	12
Forward primer	1
Reverse primer	1
gDNA (@10 ng/ul)	1
Reagent	Volume per reaction (ul)
Roche Mastermix	11
Forward primer	1
Reverse primer	1
gDNA (@10 ng/ul)	1
Dieza	1
Reagent	Volume per reaction (ul)
Roche Mastermix	11
Forward primer	1
Reverse primer	1
gDNA (@10 ng/ul)	1
DMSO	1

Table 55: PCR Mastermixes tested and utilized for all primers

### 8.1.5 MSA GWA study results

Chr	Position	Marker	Gene or nearest gene	Location	Putative function	P value	OR	Alleles**	Allele freq	R <sup>2</sup>
17	34,359,508	rs78523330	FBXO47	intronic	involved in protein ubiquitination and degradation	1.84E-07	0.45	A/G	0.96	0.55
5	60,088,977	rs7715147	ELOVL7	intronic	lipid metabolism	2.87E-07	1.47	C/A	0.74	0.68
6	12,453,679	rs16872704	EDN1	intergenic	vasoconstrictor	3.82E-07	1.51	A/G	0.84	0.94
17	41,160,977	rs9303521	CRHR1	intronic	G-protein coupled receptor, activation of signal transduction pathways	6.78E-07	0.76	T/G	0.51	0.92
			<i>MAPT*</i>	<i>Intergenic</i>	<i>microtubule binding protein</i>					

1	4,023,064	rs12044274	LOC728716	intergenic	non-coding RNA	3.30E-06	1.48	T/A	0.65	0.47
6	7,161,483	rs1413700	RREB1	intronic	zinc finger transcription factor	3.43E-06	1.39	G/C	0.79	0.92
2	239,000,140	rs473651	ASB1	intergenic	testis development	3.54E-06	0.78	C/A	0.62	0.98
17	42,218,292	rs916888	WNT3	intergenic	WNT signaling gene, embryogenesis	3.68E-06	1.35	T/C	0.75	0.96
14	47,000,222	rs78274439	MDGA2	intronic	possibly involved in cell-cell interactions	3.73E-06	1.73	A/G	0.85	0.46
20	59,576,381	rs2252187	CDH4	intronic	calcium dependent cell adhesion	3.97E-06	0.74	G/A	0.77	0.81
10	100,039,469	rs4919206	LOXL4	intergenic	biogenesis of connective tissue	4.36E-06	0.73	C/G	0.49	0.58
17	14,628,830	rs62060075	CDRT7	intergenic	non-coding RNA	4.64E-06	0.60	C/T	0.92	0.61
8	26,005,267	rs4872401	EBF2	intergenic	transcription factor	4.75E-06	0.38	G/A	0.96	0.33
5	40,890,439	rs1697938	CARD6	UTR	involved in apoptosis	4.75E-06	0.79	T/C	0.54	0.99
16	22,938,761	rs8044188	USP31	intergenic	Unknown	4.84E-06	1.38	G/A	0.79	0.95
9	128,448,334	rs10819190	LMX1B	intronic	transcription factor	5.41E-06	1.32	G/A	0.63	0.82
5	127,863,758	rs892864	FBN2	intronic	component of connective tissue microfibrils	5.96E-06	1.62	T/A	0.06	0.81
2	31,430,416	rs115903524	XDH	intronic	purine degradation	6.41E-06	0.51	G/C	0.97	0.83
19	33,413,730	rs11084877	LOC148189	intergenic	non-coding RNA	6.65E-06	0.76	A/G	0.76	0.96
2	138,023,816	rs10209086	THSD7B	intronic	Unknown	6.95E-06	0.78	T/C	0.42	0.98
9	85,989,249	rs2256039	SLC28A3	intergenic	Nucleoside transporter	7.92E-06	0.70	A/C	0.86	0.74
16	85,251,338	rs78765336	FOXL1	intergenic	Unknown	8.46E-06	0.46	G/T	0.97	0.57
7	24,585,776	rs55782418	MPP6	intronic	tumor suppression and receptor clustering	9.59E-06	0.58	G/C	0.96	0.90
17	51,515,165	rs79331640	ANKFN1	intergenic	Unknown	9.97E-06	0.45	G/A	0.97	0.54

**Table 56: MSA GWA study results**

**\*MAPT is not the closest gene in this locus but due to its known role in other neurodegenerative diseases the most likely candidate and therefore also listed here.**

\*\* first allele refers to effect allele

(Reproduced by Sailer et al. 2016)

